13 | F DISTRIBUTION AND ONE-WAY ANOVA

Figure 13.1 One-way ANOVA is used to measure information from several groups.

Introduction

Chapter Objectives

By the end of this chapter, the student should be able to:

- Interpret the *F* probability distribution as the number of groups and the sample size change.
- Discuss two uses for the *F* distribution: one-way ANOVA and the test of two variances.
- Conduct and interpret one-way ANOVA.
- Conduct and interpret hypothesis tests of two variances.

Many statistical applications in psychology, social science, business administration, and the natural sciences involve several groups. For example, an environmentalist is interested in knowing if the average amount of pollution varies in several bodies of water. A sociologist is interested in knowing if the amount of income a person earns varies according to his or her upbringing. A consumer looking for a new car might compare the average gas mileage of several models.

For hypothesis tests comparing averages between more than two groups, statisticians have developed a method called "Analysis of Variance" (abbreviated ANOVA). In this chapter, you will study the simplest form of ANOVA called single factor or one-way ANOVA. You will also study the *F* distribution, used for one-way ANOVA, and the test of two variances. This is just a very brief overview of one-way ANOVA. You will study this topic in much greater detail in future statistics courses. One-Way ANOVA, as it is presented here, relies heavily on a calculator or computer.

13.1 | One-Way ANOVA

The purpose of a one-way ANOVA test is to determine the existence of a statistically significant difference among several group means. The test actually uses **variances** to help determine if the means are equal or not. In order to perform a one-way ANOVA test, there are five basic **assumptions** to be fulfilled:

1. Each population from which a sample is taken is assumed to be normal.
2. All samples are randomly selected and independent.
3. The populations are assumed to have **equal standard deviations (or variances)**.
4. The factor is a categorical variable.
5. The response is a numerical variable.

The Null and Alternative Hypotheses

The null hypothesis is simply that all the group population means are the same. The alternative hypothesis is that at least one pair of means is different. For example, if there are *k* groups:

H_0: $\mu_1 = \mu_2 = \mu_3 = ... = \mu_k$

H_a: At least two of the group means $\mu_1, \mu_2, \mu_3, ..., \mu_k$ are not equal.

The graphs, a set of box plots representing the distribution of values with the group means indicated by a horizontal line through the box, help in the understanding of the hypothesis test. In the first graph (red box plots), H_0: $\mu_1 = \mu_2 = \mu_3$ and the three populations have the same distribution if the null hypothesis is true. The variance of the combined data is approximately the same as the variance of each of the populations.

If the null hypothesis is false, then the variance of the combined data is larger which is caused by the different means as shown in the second graph (green box plots).

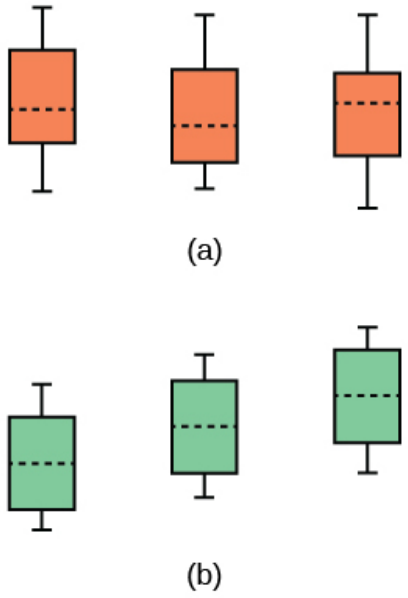

Figure 13.2 (a) H_0 is true. All means are the same; the differences are due to random variation. (b) H_0 is not true. All means are not the same; the differences are too large to be due to random variation.

13.2 | The F Distribution and the F-Ratio

The distribution used for the hypothesis test is a new one. It is called the ***F* distribution**, named after Sir Ronald Fisher, an English statistician. The *F* statistic is a ratio (a fraction). There are two sets of degrees of freedom; one for the numerator and one for the denominator.

For example, if *F* follows an *F* distribution and the number of degrees of freedom for the numerator is four, and the number of degrees of freedom for the denominator is ten, then $F \sim F_{4,10}$.

NOTE

The *F* distribution is derived from the Student's t-distribution. The values of the *F* distribution are squares of the corresponding values of the *t*-distribution. One-Way ANOVA expands the *t*-test for comparing more than two groups. The scope of that derivation is beyond the level of this course.

To calculate the ***F* ratio**, two estimates of the variance are made.

1. **Variance between samples:** An estimate of σ^2 that is the variance of the sample means multiplied by n (when the sample sizes are the same.). If the samples are different sizes, the variance between samples is weighted to account for the different sample sizes. The variance is also called **variation due to treatment or explained variation.**
2. **Variance within samples:** An estimate of σ^2 that is the average of the sample variances (also known as a pooled variance). When the sample sizes are different, the variance within samples is weighted. The variance is also called the **variation due to error or unexplained variation.**

- $SS_{between}$ = the **sum of squares** that represents the variation among the different samples
- SS_{within} = the sum of squares that represents the variation within samples that is due to chance.

To find a "sum of squares" means to add together squared quantities that, in some cases, may be weighted. We used sum of squares to calculate the sample variance and the sample standard deviation in **Descriptive Statistics**.

MS means " **mean square**." $MS_{between}$ is the variance between groups, and MS_{within} is the variance within groups.

Calculation of Sum of Squares and Mean Square

- k = the number of different groups
- n_j = the size of the j^{th} group
- s_j = the sum of the values in the j^{th} group
- n = total number of all the values combined (total sample size: $\sum n_j$)
- x = one value: $\sum x = \sum s_j$
- Sum of squares of all values from every group combined: $\sum x^2$
- Between group variability: $SS_{total} = \sum x^2 - \frac{\left(\sum x^2\right)}{n}$
- Total sum of squares: $\sum x^2 - \frac{(\sum x)^2}{n}$
- Explained variation: sum of squares representing variation among the different samples: $SS_{between} = \sum\left[\frac{(sj)^2}{n_j}\right] - \frac{(\sum s_j)^2}{n}$
- Unexplained variation: sum of squares representing variation within samples due to chance: $SS_{within} = SS_{total} - SS_{between}$
- *df*'s for different groups (*df*'s for the numerator): $df = k - 1$
- Equation for errors within samples (*df*'s for the denominator): $df_{within} = n - k$
- Mean square (variance estimate) explained by the different groups: $MS_{between} = \frac{SS_{between}}{df_{between}}$

- Mean square (variance estimate) that is due to chance (unexplained): $MS_{\text{within}} = \frac{SS_{\text{within}}}{df_{\text{within}}}$

MS_{between} and MS_{within} can be written as follows:

- $MS_{\text{between}} = \frac{SS_{\text{between}}}{df_{\text{between}}} = \frac{SS_{\text{between}}}{k-1}$
- $MS_{within} = \frac{SS_{within}}{df_{within}} = \frac{SS_{within}}{n-k}$

The one-way ANOVA test depends on the fact that MS_{between} can be influenced by population differences among means of the several groups. Since MS_{within} compares values of each group to its own group mean, the fact that group means might be different does not affect MS_{within}.

The null hypothesis says that all groups are samples from populations having the same normal distribution. The alternate hypothesis says that at least two of the sample groups come from populations with different normal distributions. If the null hypothesis is true, MS_{between} and MS_{within} should both estimate the same value.

NOTE

The null hypothesis says that all the group population means are equal. The hypothesis of equal means implies that the populations have the same normal distribution, because it is assumed that the populations are normal and that they have equal variances.

F-Ratio or *F* Statistic

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}}$$

If MS_{between} and MS_{within} estimate the same value (following the belief that H_0 is true), then the F-ratio should be approximately equal to one. Mostly, just sampling errors would contribute to variations away from one. As it turns out, MS_{between} consists of the population variance plus a variance produced from the differences between the samples. MS_{within} is an estimate of the population variance. Since variances are always positive, if the null hypothesis is false, MS_{between} will generally be larger than MS_{within}.Then the F-ratio will be larger than one. However, if the population effect is small, it is not unlikely that MS_{within} will be larger in a given sample.

The foregoing calculations were done with groups of different sizes. If the groups are the same size, the calculations simplify somewhat and the F-ratio can be written as:

F-Ratio Formula when the groups are the same size

$$F = \frac{n \cdot s_{\bar{x}}^2}{s^2_{\text{pooled}}}$$

where ...

- n = the sample size
- $df_{\text{numerator}} = k - 1$
- $df_{\text{denominator}} = n - k$
- s^2 pooled = the mean of the sample variances (pooled variance)
- $s_{\bar{x}}^2$ = the variance of the sample means

Data are typically put into a table for easy viewing. One-Way ANOVA results are often displayed in this manner by computer software.

Source of Variation	Sum of Squares (*SS*)	Degrees of Freedom (*df*)	Mean Square (*MS*)	*F*
Factor (Between)	*SS*(Factor)	$k - 1$	*MS*(Factor) = *SS*(Factor)/($k - 1$)	*F* = *MS*(Factor)/*MS*(Error)
Error (Within)	*SS*(Error)	$n - k$	*MS*(Error) = *SS*(Error)/($n - k$)	
Total	*SS*(Total)	$n - 1$		

Table 13.1

Example 13.1

Three different diet plans are to be tested for mean weight loss. The entries in the table are the weight losses for the different plans. The one-way ANOVA results are shown in **Table 13.2**.

Plan 1: $n_1 = 4$	Plan 2: $n_2 = 3$	Plan 3: $n_3 = 3$
5	3.5	8
4.5	7	4
4		3.5
3	4.5	

Table 13.2

$s_1 = 16.5$, $s_2 = 15$, $s_3 = 15.7$

Following are the calculations needed to fill in the one-way ANOVA table. The table is used to conduct a hypothesis test.

$$SS(between) = \sum \left[\frac{(s_j)^2}{n_j} \right] - \frac{\left(\sum s_j\right)^2}{n}$$

$$= \frac{s_1^2}{4} + \frac{s_2^2}{3} + \frac{s_3^2}{3} - \frac{(s_1 + s_2 + s_3)^2}{10}$$

where $n_1 = 4$, $n_2 = 3$, $n_3 = 3$ and $n = n_1 + n_2 + n_3 = 10$

$$= \frac{(16.5)^2}{4} + \frac{(15)^2}{3} + \frac{(5.5)^2}{3} - \frac{(16.5 + 15 + 15.5)^2}{10}$$

$$SS(between) = 2.2458$$

$$S(total) = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$= \left(5^2 + 4.5^2 + 4^2 + 3^2 + 3.5^2 + 7^2 + 4.5^2 + 8^2 + 4^2 + 3.5^2\right)$$

$$- \frac{(5 + 4.5 + 4 + 3 + 3.5 + 7 + 4.5 + 8 + 4 + 3.5)^2}{10}$$

$$= 244 - \frac{47^2}{10} = 244 - 220.9$$

$$SS(total) = 23.1$$

$$SS(within) = SS(total) - SS(between)$$

$$= 23.1 - 2.2458$$

$$SS(within) = 20.8542$$

Using the TI-83, 83+, 84, 84+ Calculator

One-Way ANOVA Table: The formulas for *SS*(Total), *SS*(Factor) = *SS*(Between) and *SS*(Error) = *SS*(Within) as shown previously. The same information is provided by the TI calculator hypothesis test function ANOVA in STAT TESTS (syntax is ANOVA(L1, L2, L3) where L1, L2, L3 have the data from Plan 1, Plan 2, Plan 3 respectively).

Source of Variation	**Sum of Squares (*SS*)**	**Degrees of Freedom (*df*)**	**Mean Square (*MS*)**	***F***
Factor (Between)	*SS*(Factor) = *SS*(Between) = 2.2458	$k - 1$ = 3 groups – 1 = 2	*MS*(Factor) = *SS*(Factor)/(k – 1) = 2.2458/2 = 1.1229	F = *MS*(Factor)/*MS*(Error) = 1.1229/2.9792 = 0.3769
Error (Within)	*SS*(Error) = *SS*(Within) = 20.8542	$n - k$ = 10 total data – 3 groups = 7	*MS*(Error) = *SS*(Error)/(n – k) = 20.8542/7 = 2.9792	
Total	*SS*(Total) = 2.2458 + 20.8542 = 23.1	$n - 1$ = 10 total data – 1 = 9		

Table 13.3

Try It

13.1 As part of an experiment to see how different types of soil cover would affect slicing tomato production, Marist College students grew tomato plants under different soil cover conditions. Groups of three plants each had one of the following treatments

- bare soil
- a commercial ground cover
- black plastic
- straw
- compost

All plants grew under the same conditions and were the same variety. Students recorded the weight (in grams) of tomatoes produced by each of the $n = 15$ plants:

Bare: $n_1 = 3$	**Ground Cover: $n_2 = 3$**	**Plastic: $n_3 = 3$**	**Straw: $n_4 = 3$**	**Compost: $n_5 = 3$**
2,625	5,348	6,583	7,285	6,277
2,997	5,682	8,560	6,897	7,818
4,915	5,482	3,830	9,230	8,677

Table 13.4

Create the one-way ANOVA table.

The one-way ANOVA hypothesis test is always right-tailed because larger F-values are way out in the right tail of the F-distribution curve and tend to make us reject H_0.

Notation

The notation for the F distribution is $F \sim F_{df(num),df(denom)}$

where $df(num) = df_{\text{between}}$ and $df(denom) = df_{\text{within}}$

The mean for the F distribution is $\mu = \frac{df(num)}{df(denom) - 1}$

13.3 | Facts About the F Distribution

Here are some facts about the *F* distribution.

1. The curve is not symmetrical but skewed to the right.
2. There is a different curve for each set of *df*s.
3. The F statistic is greater than or equal to zero.
4. As the degrees of freedom for the numerator and for the denominator get larger, the curve approximates the normal.
5. Other uses for the F distribution include comparing two variances and two-way Analysis of Variance. Two-Way Analysis is beyond the scope of this chapter.

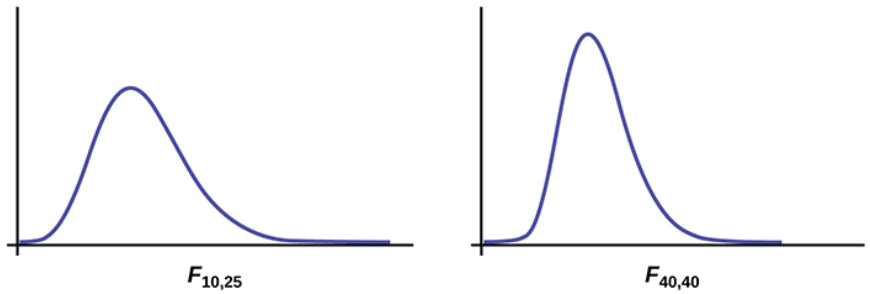

Figure 13.3

Example 13.2

Let's return to the slicing tomato exercise in **Try It**. The means of the tomato yields under the five mulching conditions are represented by μ_1, μ_2, μ_3, μ_4, μ_5. We will conduct a hypothesis test to determine if all means are the same or at least one is different. Using a significance level of 5%, test the null hypothesis that there is no difference in mean yields among the five groups against the alternative hypothesis that at least one mean is different from the rest.

Solution 13.2

The null and alternative hypotheses are:

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

H_a: $\mu_i \neq \mu_j$ some $i \neq j$

The one-way ANOVA results are shown in **Table 13.4**

Source of Variation	Sum of Squares (*SS*)	Degrees of Freedom (*df*)	Mean Square (*MS*)	*F*
Factor (Between)	36,648,561	5 – 1 = 4	$\frac{36,648,561}{4} = 9,162,140$	$\frac{9,162,140}{2,044,672.6} = 4.4810$
Error (Within)	20,446,726	15 – 5 = 10	$\frac{20,446,726}{10} = 2,044,672.6$	
Total	57,095,287	15 – 1 = 14		

Table 13.5

Distribution for the test: $F_{4,10}$

$df(num) = 5 – 1 = 4$

$df(denom) = 15 – 5 = 10$

Test statistic: $F = 4.4810$

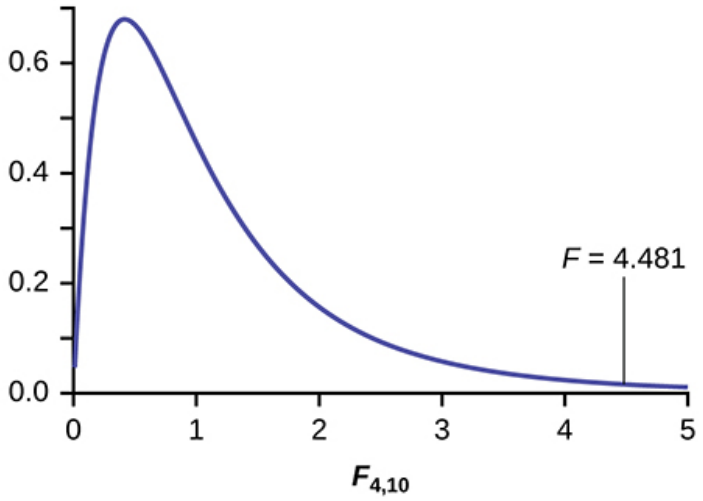

Figure 13.4

Probability Statement: p-value = $P(F > 4.481) = 0.0248$.

Compare α and the p-value: $\alpha = 0.05$, p-value = 0.0248

Make a decision: Since $\alpha > p$-value, we reject H_0.

Conclusion: At the 5% significance level, we have reasonably strong evidence that differences in mean yields for slicing tomato plants grown under different mulching conditions are unlikely to be due to chance alone. We may conclude that at least some of mulches led to different mean yields.

Using the TI-83, 83+, 84, 84+ Calculator

To find these results on the calculator:

Press STAT. Press 1:EDIT. Put the data into the lists L_1, L_2, L_3, L_4, L_5.

Press STAT, and arrow over to TESTS, and arrow down to ANOVA. Press ENTER, and then enter L_1, L_2, L_3, L_4, L_5). Press ENTER. You will see that the values in the foregoing ANOVA table are easily produced by the calculator, including the test statistic and the p-value of the test.

The calculator displays:
$F = 4.4810$
$p = 0.0248$ (p-value)
Factor
$df = 4$
$SS = 36648560.9$
$MS = 9162140.23$
Error
$df = 10$
$SS = 20446726$
$MS = 2044672.6$

Try It

13.2 MRSA, or *Staphylococcus aureus*, can cause a serious bacterial infections in hospital patients. **Table 13.6** shows various colony counts from different patients who may or may not have MRSA.

Conc = 0.6	Conc = 0.8	Conc = 1.0	Conc = 1.2	Conc = 1.4
9	16	22	30	27
66	93	147	199	168
98	82	120	148	132

Table 13.6

Plot of the data for the different concentrations:

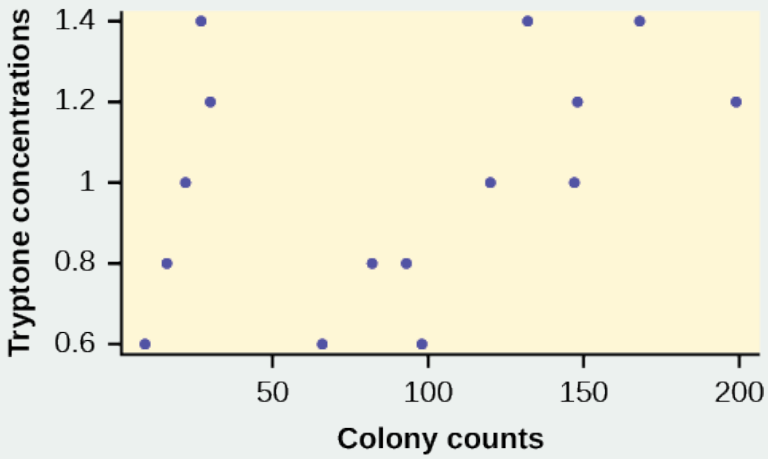

Figure 13.5

Test whether the mean number of colonies are the same or are different. Construct the ANOVA table (by hand or by using a TI-83, 83+, or 84+ calculator), find the p-value, and state your conclusion. Use a 5% significance level.

Example 13.3

Four sororities took a random sample of sisters regarding their grade means for the past term. The results are shown in **Table 13.7**.

Sorority 1	Sorority 2	Sorority 3	Sorority 4
2.17	2.63	2.63	3.79
1.85	1.77	3.78	3.45
2.83	3.25	4.00	3.08
1.69	1.86	2.55	2.26
3.33	2.21	2.45	3.18

Table 13.7 MEAN GRADES FOR FOUR SORORITIES

Using a significance level of 1%, is there a difference in mean grades among the sororities?

Solution 13.3

Let μ_1, μ_2, μ_3, μ_4 be the population means of the sororities. Remember that the null hypothesis claims that the sorority groups are from the same normal distribution. The alternate hypothesis says that at least two of the sorority groups come from populations with different normal distributions. Notice that the four sample sizes are each five.

NOTE

This is an example of a **balanced design**, because each factor (i.e., sorority) has the same number of observations.

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

H_a: Not all of the means μ_1, μ_2, μ_3, μ_4 are equal.

Distribution for the test: $F_{3,16}$

where $k = 4$ groups and $n = 20$ samples in total

$df(num) = k - 1 = 4 - 1 = 3$

$df(denom) = n - k = 20 - 4 = 16$

Calculate the test statistic: $F = 2.23$

Graph:

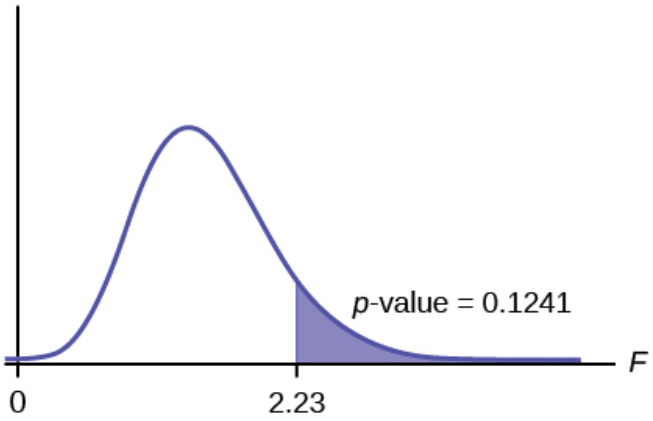

Figure 13.6

Probability statement: p-value = $P(F > 2.23) = 0.1241$

Compare α and the p-value: $\alpha = 0.01$
p-value = 0.1241
$\alpha < p$-value

Make a decision: Since $\alpha < p$-value, you cannot reject H_0.

Conclusion: There is not sufficient evidence to conclude that there is a difference among the mean grades for the sororities.

Using the TI-83, 83+, 84, 84+ Calculator

Put the data into lists L_1, L_2, L_3, and L_4. Press STAT and arrow over to TESTS. Arrow down to F:ANOVA. Press ENTER and Enter (L1,L2,L3,L4).

The calculator displays the F statistic, the p-value and the values for the one-way ANOVA table:
$F = 2.2303$
$p = 0.1241$ (p-value)
Factor
$df = 3$
$SS = 2.88732$
$MS = 0.96244$
Error
$df = 16$
$SS = 6.9044$
$MS = 0.431525$

Try It

13.3 Four sports teams took a random sample of players regarding their GPAs for the last year. The results are shown in **Table 13.8**.

Basketball	Baseball	Hockey	Lacrosse
3.6	2.1	4.0	2.0
2.9	2.6	2.0	3.6
2.5	3.9	2.6	3.9
3.3	3.1	3.2	2.7
3.8	3.4	3.2	2.5

Table 13.8 GPAs FOR FOUR SPORTS TEAMS

Use a significance level of 5%, and determine if there is a difference in GPA among the teams.

Example 13.4

A fourth grade class is studying the environment. One of the assignments is to grow bean plants in different soils. Tommy chose to grow his bean plants in soil found outside his classroom mixed with dryer lint. Tara chose to grow her bean plants in potting soil bought at the local nursery. Nick chose to grow his bean plants in soil from his mother's garden. No chemicals were used on the plants, only water. They were grown inside the classroom next to a large window. Each child grew five plants. At the end of the growing period, each plant was measured, producing the data (in inches) in **Table 13.9**.

Tommy's Plants	Tara's Plants	Nick's Plants
24	25	23
21	31	27
23	23	22
30	20	30
23	28	20

Table 13.9

Does it appear that the three media in which the bean plants were grown produce the same mean height? Test at a 3% level of significance.

Solution 13.4

This time, we will perform the calculations that lead to the F' statistic. Notice that each group has the same number of plants, so we will use the formula $F' = \frac{n \cdot s_{\bar{x}}^2}{s^2_{\text{pooled}}}$.

First, calculate the sample mean and sample variance of each group.

	Tommy's Plants	Tara's Plants	Nick's Plants
Sample Mean	24.2	25.4	24.4
Sample Variance	11.7	18.3	16.3

Table 13.10

Next, calculate the variance of the three group means (Calculate the variance of 24.2, 25.4, and 24.4). **Variance of the group means = 0.413** = $s_{\bar{x}}^2$

Then $MS_{between} = ns_{\bar{x}}^2 = (5)(0.413)$ where $n = 5$ is the sample size (number of plants each child grew).

Calculate the mean of the three sample variances (Calculate the mean of 11.7, 18.3, and 16.3). **Mean of the sample variances = 15.433** = s^2 ***pooled***

Then $MS_{within} = s^2_{pooled} = 15.433$.

The F statistic (or F ratio) is $F = \frac{MS_{\text{between}}}{MS_{\text{within}}} = \frac{ns_{\bar{x}}^2}{s^2_{pooled}} = \frac{(5)(0.413)}{15.433} = 0.134$

The *dfs* for the numerator = the number of groups – 1 = 3 – 1 = 2.

The *dfs* for the denominator = the total number of samples – the number of groups = 15 – 3 = 12

The distribution for the test is $F_{2,12}$ and the F statistic is $F = 0.134$

The p-value is $P(F > 0.134) = 0.8759$.

Decision: Since $\alpha = 0.03$ and the p-value = 0.8759, do not reject H_0. (Why?)

Conclusion: With a 3% level of significance, from the sample data, the evidence is not sufficient to conclude that the mean heights of the bean plants are different.

Using the TI-83, 83+, 84, 84+ Calculator

To calculate the p-value:

*Press 2nd DISTR

*Arrow down to Fcdf(and press ENTER.

*Enter 0.134, E99, 2, 12)

*Press ENTER

The p-value is 0.8759.

Try It

13.4 Another fourth grader also grew bean plants, but this time in a jelly-like mass. The heights were (in inches) 24, 28, 25, 30, and 32. Do a one-way ANOVA test on the four groups. Are the heights of the bean plants different? Use the same method as shown in **Example 13.4**.

Collaborative Exercise

From the class, create four groups of the same size as follows: men under 22, men at least 22, women under 22, women at least 22. Have each member of each group record the number of states in the United States he or she has visited. Run an ANOVA test to determine if the average number of states visited in the four groups are the same. Test at a 1% level of significance. Use one of the solution sheets in **Appendix E**.

13.4 | Test of Two Variances

Another of the uses of the F distribution is testing two variances. It is often desirable to compare two variances rather than two averages. For instance, college administrators would like two college professors grading exams to have the same variation in their grading. In order for a lid to fit a container, the variation in the lid and the container should be the same. A supermarket might be interested in the variability of check-out times for two checkers.

In order to perform a F test of two variances, it is important that the following are true:

1. The populations from which the two samples are drawn are normally distributed.
2. The two populations are independent of each other.

Unlike most other tests in this book, the F test for equality of two variances is very sensitive to deviations from normality. If the two distributions are not normal, the test can give higher p-values than it should, or lower ones, in ways that are unpredictable. Many texts suggest that students not use this test at all, but in the interest of completeness we include it here.

Suppose we sample randomly from two independent normal populations. Let σ_1^2 and σ_2^2 be the population variances and s_1^2 and s_2^2 be the sample variances. Let the sample sizes be n_1 and n_2. Since we are interested in comparing the two sample variances, we use the F ratio:

$$F = \frac{\left[\frac{(s_1)^2}{(\sigma_1)^2}\right]}{\left[\frac{(s_2)^2}{(\sigma_2)^2}\right]}$$

F has the distribution $F \sim F(n_1 - 1, n_2 - 1)$

where $n_1 - 1$ are the degrees of freedom for the numerator and $n_2 - 1$ are the degrees of freedom for the denominator.

If the null hypothesis is $\sigma_1^2 = \sigma_2^2$, then the F Ratio becomes $F = \frac{\left[\frac{(s_1)^2}{(\sigma_1)^2}\right]}{\left[\frac{(s_2)^2}{(\sigma_2)^2}\right]} = \frac{(s_1)^2}{(s_2)^2}$.

NOTE

The F ratio could also be $\frac{(s_2)^2}{(s_1)^2}$. It depends on H_a and on which sample variance is larger.

If the two populations have equal variances, then s_1^2 and s_2^2 are close in value and $F = \frac{(s_1)^2}{(s_2)^2}$ is close to one. But if the two population variances are very different, s_1^2 and s_2^2 tend to be very different, too. Choosing s_1^2 as the larger sample variance causes the ratio $\frac{(s_1)^2}{(s_2)^2}$ to be greater than one. If s_1^2 and s_2^2 are far apart, then $F = \frac{(s_1)^2}{(s_2)^2}$ is a large number.

Therefore, if F is close to one, the evidence favors the null hypothesis (the two population variances are equal). But if F is much larger than one, then the evidence is against the null hypothesis. **A test of two variances may be left, right, or two-tailed.**

Example 13.5

Two college instructors are interested in whether or not there is any variation in the way they grade math exams. They each grade the same set of 30 exams. The first instructor's grades have a variance of 52.3. The second instructor's grades have a variance of 89.9. Test the claim that the first instructor's variance is smaller. (In most colleges, it is desirable for the variances of exam grades to be nearly the same among instructors.) The level of significance is 10%.

Solution 13.5

Let 1 and 2 be the subscripts that indicate the first and second instructor, respectively.

$n_1 = n_2 = 30$.

H_0: $\sigma_1^2 = \sigma_2^2$ and H_a: $\sigma_1^2 < \sigma_2^2$

Calculate the test statistic: By the null hypothesis $(\sigma_1^2 = \sigma_2^2)$, the F statistic is:

$$F = \frac{\left[\frac{(s_1)^2}{(\sigma_1)^2}\right]}{\left[\frac{(s_2)^2}{(\sigma_2)^2}\right]} = \frac{(s_1)^2}{(s_2)^2} = \frac{52.3}{89.9} = 0.5818$$

Distribution for the test: $F_{29,29}$ where $n_1 - 1 = 29$ and $n_2 - 1 = 29$.

Graph: This test is left tailed.

Draw the graph labeling and shading appropriately.

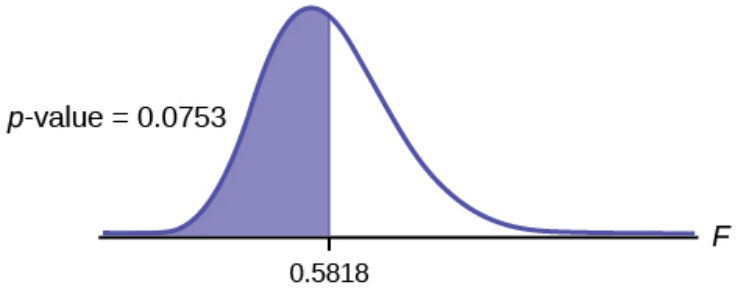

Figure 13.7

Probability statement: p-value = $P(F < 0.5818) = 0.0753$

Compare α and the p-value: $\alpha = 0.10$ $\alpha > p$-value.

Make a decision: Since $\alpha > p$-value, reject H_0.

Conclusion: With a 10% level of significance, from the data, there is sufficient evidence to conclude that the variance in grades for the first instructor is smaller.

Using the TI-83, 83+, 84, 84+ Calculator

Press STAT and arrow over to TESTS. Arrow down to D:2-SampFTest. Press ENTER. Arrow to Stats and press ENTER. For Sx1, n1, Sx2, and n2, enter $\sqrt{(52.3)}$, 30, $\sqrt{(89.9)}$, and 30. Press ENTER after each. Arrow to σ1: and < σ2. Press ENTER. Arrow down to Calculate and press ENTER. $F = 0.5818$ and p-value = 0.0753. Do the procedure again and try Draw instead of Calculate.

Try It Σ

13.5 The New York Choral Society divides male singers up into four categories from highest voices to lowest: Tenor1, Tenor2, Bass1, Bass2. In the table are heights of the men in the Tenor1 and Bass2 groups. One suspects that taller men will have lower voices, and that the variance of height may go up with the lower voices as well. Do we have good evidence that the variance of the heights of singers in each of these two groups (Tenor1 and Bass2) are different?

Tenor1	Bass2	Tenor 1	Bass 2	Tenor 1	Bass 2
69	72	67	72	68	67
72	75	70	74	67	70
71	67	65	70	64	70
66	75	72	66		69
76	74	70	68		72
74	72	68	75		71
71	72	64	68		74
66	74	73	70		75
68	72	66	72		

Table 13.11

13.5 | Lab: One-Way ANOVA

Stats Lab

13.1 One-Way ANOVA

Class Time:

Names:

Student Learning Outcome

- The student will conduct a simple one-way ANOVA test involving three variables.

Collect the Data

1. Record the price per pound of eight fruits, eight vegetables, and eight breads in your local supermarket.

Fruits	Vegetables	Breads

Table 13.12

2. Explain how you could try to collect the data randomly.

Analyze the Data and Conduct a Hypothesis Test

1. Compute the following:
   a. Fruit:
      i. $\bar{x}$ = _____
      ii. s_x = _____
      iii. n = _____
   b. Vegetables:
      i. $\bar{x}$ = _____
      ii. s_x = _____
      iii. n = _____
   c. Bread:
      i. $\bar{x}$ = _____
      ii. s_x = _____
      iii. n = _____
2. Find the following:
   a. $df(num)$ = _____
   b. $df(denom)$ = _____

3. State the approximate distribution for the test.
4. Test statistic: $F =$ _____
5. Sketch a graph of this situation. CLEARLY, label and scale the horizontal axis and shade the region(s) corresponding to the *p*-value.
6. *p*-value = _____
7. Test at $\alpha = 0.05$. State your decision and conclusion.
8. a. Decision: Why did you make this decision?
   b. Conclusion (write a complete sentence).
   c. Based on the results of your study, is there a need to investigate any of the food groups' prices? Why or why not?

KEY TERMS

Analysis of Variance also referred to as ANOVA, is a method of testing whether or not the means of three or more populations are equal. The method is applicable if:

- all populations of interest are normally distributed.
- the populations have equal standard deviations.
- samples (not necessarily of the same size) are randomly and independently selected from each population.

The test statistic for analysis of variance is the *F*-ratio.

One-Way ANOVA a method of testing whether or not the means of three or more populations are equal; the method is applicable if:

- all populations of interest are normally distributed.
- the populations have equal standard deviations.
- samples (not necessarily of the same size) are randomly and independently selected from each population.

The test statistic for analysis of variance is the *F*-ratio.

Variance mean of the squared deviations from the mean; the square of the standard deviation. For a set of data, a deviation can be represented as $x - \bar{x}$ where x is a value of the data and $\bar{x}$ is the sample mean. The sample variance is equal to the sum of the squares of the deviations divided by the difference of the sample size and one.

CHAPTER REVIEW

13.1 One-Way ANOVA

Analysis of variance extends the comparison of two groups to several, each a level of a categorical variable (factor). Samples from each group are independent, and must be randomly selected from normal populations with equal variances. We test the null hypothesis of equal means of the response in every group versus the alternative hypothesis of one or more group means being different from the others. A one-way ANOVA hypothesis test determines if several population means are equal. The distribution for the test is the *F* distribution with two different degrees of freedom.

Assumptions:

1. Each population from which a sample is taken is assumed to be normal.
2. All samples are randomly selected and independent.
3. The populations are assumed to have equal standard deviations (or variances).

13.2 The F Distribution and the F-Ratio

Analysis of variance compares the means of a response variable for several groups. ANOVA compares the variation within each group to the variation of the mean of each group. The ratio of these two is the *F* statistic from an *F* distribution with (number of groups – 1) as the numerator degrees of freedom and (number of observations – number of groups) as the denominator degrees of freedom. These statistics are summarized in the ANOVA table.

13.3 Facts About the F Distribution

The graph of the *F* distribution is always positive and skewed right, though the shape can be mounded or exponential depending on the combination of numerator and denominator degrees of freedom. The *F* statistic is the ratio of a measure of the variation in the group means to a similar measure of the variation within the groups. If the null hypothesis is correct, then the numerator should be small compared to the denominator. A small *F* statistic will result, and the area under the *F* curve to the right will be large, representing a large *p*-value. When the null hypothesis of equal group means is incorrect, then the numerator should be large compared to the denominator, giving a large *F* statistic and a small area (small *p*-value) to the right of the statistic under the *F* curve.

When the data have unequal group sizes (unbalanced data), then techniques from **Section 13.2** need to be used for hand calculations. In the case of balanced data (the groups are the same size) however, simplified calculations based on group

means and variances may be used. In practice, of course, software is usually employed in the analysis. As in any analysis, graphs of various sorts should be used in conjunction with numerical techniques. Always look of your data!

13.4 Test of Two Variances

The F test for the equality of two variances rests heavily on the assumption of normal distributions. The test is unreliable if this assumption is not met. If both distributions are normal, then the ratio of the two sample variances is distributed as an F statistic, with numerator and denominator degrees of freedom that are one less than the samples sizes of the corresponding two groups. A **test of two variances** hypothesis test determines if two variances are the same. The distribution for the hypothesis test is the F distribution with two different degrees of freedom.

Assumptions:

1. The populations from which the two samples are drawn are normally distributed.
2. The two populations are independent of each other.

FORMULA REVIEW

13.2 The F Distribution and the F-Ratio

$$SS_{\text{between}} = \sum \left[\frac{(s_j)^2}{n_j} \right] - \frac{\left(\sum s_j\right)^2}{n}$$

$$SS_{\text{total}} = \sum x^2 - \frac{\left(\sum x\right)^2}{n}$$

$$SS_{\text{within}} = SS_{\text{total}} - SS_{\text{between}}$$

$df_{\text{between}} = df(num) = k - 1$

$df_{\text{within}} = df(denom) = n - k$

$$MS_{\text{between}} = \frac{SS_{\text{between}}}{df_{\text{between}}}$$

$$MS_{\text{within}} = \frac{SS_{\text{within}}}{df_{\text{within}}}$$

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}}$$

F ratio when the groups are the same size: $F = \dfrac{n s_{\bar{x}}^2}{s^2_{pooled}}$

Mean of the F distribution: $\mu = \dfrac{df(num)}{df(denom) - 1}$

where:

- k = the number of groups
- n_j = the size of the j^{th} group
- s_j = the sum of the values in the j^{th} group
- n = the total number of all values (observations) combined
- x = one value (one observation) from the data
- $s_{\bar{x}}^2$ = the variance of the sample means
- s^2_{pooled} = the mean of the sample variances (pooled variance)

13.4 Test of Two Variances

F has the distribution $F \sim F(n_1 - 1, n_2 - 1)$

$$F = \frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}}$$

If $\sigma_1 = \sigma_2$, then $F = \dfrac{s_1^2}{s_2^2}$

PRACTICE

13.1 One-Way ANOVA

Use the following information to answer the next five exercises. There are five basic assumptions that must be fulfilled in order to perform a one-way ANOVA test. What are they?

1. Write one assumption.

2. Write another assumption.

3. Write a third assumption.

4. Write a fourth assumption.

5. Write the final assumption.

6. State the null hypothesis for a one-way ANOVA test if there are four groups.

7. State the alternative hypothesis for a one-way ANOVA test if there are three groups.

8. When do you use an ANOVA test?

13.2 The F Distribution and the F-Ratio

Use the following information to answer the next eight exercises. Groups of men from three different areas of the country are to be tested for mean weight. The entries in the table are the weights for the different groups. The one-way ANOVA results are shown in **Table 13.13**.

Group 1	Group 2	Group 3
216	202	170
198	213	165
240	284	182
187	228	197
176	210	201

Table 13.13

9. What is the Sum of Squares Factor?

10. What is the Sum of Squares Error?

11. What is the *df* for the numerator?

12. What is the *df* for the denominator?

13. What is the Mean Square Factor?

14. What is the Mean Square Error?

15. What is the *F* statistic?

Use the following information to answer the next eight exercises. Girls from four different soccer teams are to be tested for mean goals scored per game. The entries in the table are the goals per game for the different teams. The one-way ANOVA results are shown in **Table 13.14**.

Team 1	Team 2	Team 3	Team 4
1	2	0	3
2	3	1	4
0	2	1	4
3	4	0	3
2	4	0	2

Table 13.14

16. What is $SS_{between}$?

17. What is the *df* for the numerator?

18. What is $MS_{between}$?

19. What is SS_{within}?

20. What is the *df* for the denominator?

21. What is MS_{within}?

22. What is the *F* statistic?

23. Judging by the *F* statistic, do you think it is likely or unlikely that you will reject the null hypothesis?

13.3 Facts About the F Distribution

24. An *F* statistic can have what values?

25. What happens to the curves as the degrees of freedom for the numerator and the denominator get larger?

Use the following information to answer the next seven exercise. Four basketball teams took a random sample of players regarding how high each player can jump (in inches). The results are shown in **Table 13.15**.

Team 1	Team 2	Team 3	Team 4	Team 5
36	32	48	38	41
42	35	50	44	39
51	38	39	46	40

Table 13.15

26. What is the *df(num)*?

27. What is the *df(denom)*?

28. What are the Sum of Squares and Mean Squares Factors?

29. What are the Sum of Squares and Mean Squares Errors?

30. What is the *F* statistic?

31. What is the *p*-value?

32. At the 5% significance level, is there a difference in the mean jump heights among the teams?

Use the following information to answer the next seven exercises. A video game developer is testing a new game on three different groups. Each group represents a different target market for the game. The developer collects scores from a random sample from each group. The results are shown in **Table 13.16**

Group A	Group B	Group C
101	151	101
108	149	109
98	160	198
107	112	186
111	126	160

Table 13.16

33. What is the *df(num)*?

34. What is the *df(denom)*?

35. What are the $SS_{between}$ and $MS_{between}$?

36. What are the SS_{within} and MS_{within}?

37. What is the *F* Statistic?

38. What is the *p*-value?

39. At the 10% significance level, are the scores among the different groups different?

Use the following information to answer the next three exercises. Suppose a group is interested in determining whether teenagers obtain their drivers licenses at approximately the same average age across the country. Suppose that the following

data are randomly collected from five teenagers in each region of the country. The numbers represent the age at which teenagers obtained their drivers licenses.

	Northeast	South	West	Central	East
	16.3	16.9	16.4	16.2	17.1
	16.1	16.5	16.5	16.6	17.2
	16.4	16.4	16.6	16.5	16.6
	16.5	16.2	16.1	16.4	16.8
$\bar{x} =$	_____	_____	_____	_____	_____
$s^2 =$	_____	_____	_____	_____	_____

Table 13.17

Enter the data into your calculator or computer.

40. p-value = _____

State the decisions and conclusions (in complete sentences) for the following preconceived levels of α.

41. $\alpha = 0.05$

a. Decision: _____

b. Conclusion: _____

42. $\alpha = 0.01$

a. Decision: _____

b. Conclusion: _____

13.4 Test of Two Variances

Use the following information to answer the next two exercises. There are two assumptions that must be true in order to perform an F test of two variances.

43. Name one assumption that must be true.

44. What is the other assumption that must be true?

Use the following information to answer the next five exercises. Two coworkers commute from the same building. They are interested in whether or not there is any variation in the time it takes them to drive to work. They each record their times for 20 commutes. The first worker's times have a variance of 12.1. The second worker's times have a variance of 16.9. The first worker thinks that he is more consistent with his commute times and that his commute time is shorter. Test the claim at the 10% level.

45. State the null and alternative hypotheses.

46. What is s_1 in this problem?

47. What is s_2 in this problem?

48. What is n?

49. What is the F statistic?

50. What is the p-value?

51. Is the claim accurate?

Use the following information to answer the next four exercises. Two students are interested in whether or not there is variation in their test scores for math class. There are 15 total math tests they have taken so far. The first student's grades have a standard deviation of 38.1. The second student's grades have a standard deviation of 22.5. The second student thinks his scores are lower.

52. State the null and alternative hypotheses.

53. What is the *F* Statistic?

54. What is the *p*-value?

55. At the 5% significance level, do we reject the null hypothesis?

Use the following information to answer the next three exercises. Two cyclists are comparing the variances of their overall paces going uphill. Each cyclist records his or her speeds going up 35 hills. The first cyclist has a variance of 23.8 and the second cyclist has a variance of 32.1. The cyclists want to see if their variances are the same or different.

56. State the null and alternative hypotheses.

57. What is the *F* Statistic?

58. At the 5% significance level, what can we say about the cyclists' variances?

HOMEWORK

13.1 One-Way ANOVA

59. Three different traffic routes are tested for mean driving time. The entries in the table are the driving times in minutes on the three different routes. The one-way ANOVA results are shown in **Table 13.18**.

Route 1	Route 2	Route 3
30	27	16
32	29	41
27	28	22
35	36	31

Table 13.18

State $SS_{between}$, SS_{within}, and the *F* statistic.

60. Suppose a group is interested in determining whether teenagers obtain their drivers licenses at approximately the same average age across the country. Suppose that the following data are randomly collected from five teenagers in each region of the country. The numbers represent the age at which teenagers obtained their drivers licenses.

	Northeast	South	West	Central	East
	16.3	16.9	16.4	16.2	17.1
	16.1	16.5	16.5	16.6	17.2
	16.4	16.4	16.6	16.5	16.6
	16.5	16.2	16.1	16.4	16.8
$\bar{x} =$	_____	_____	_____	_____	_____
$s^2 =$	_____	_____	_____	_____	_____

Table 13.19

State the hypotheses.

H_0: _____

H_a: _____

13.2 The F Distribution and the F-Ratio

Use the following information to answer the next three exercises. Suppose a group is interested in determining whether teenagers obtain their drivers licenses at approximately the same average age across the country. Suppose that the following

data are randomly collected from five teenagers in each region of the country. The numbers represent the age at which teenagers obtained their drivers licenses.

	Northeast	South	West	Central	East
	16.3	16.9	16.4	16.2	17.1
	16.1	16.5	16.5	16.6	17.2
	16.4	16.4	16.6	16.5	16.6
	16.5	16.2	16.1	16.4	16.8
$\bar{x} =$	_____	_____	_____	_____	_____
$s^2 =$	_____	_____	_____	_____	_____

Table 13.20

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

$H\alpha$: At least any two of the group means $\mu_1, \mu_2, \ldots, \mu_5$ are not equal.

61. degrees of freedom – numerator: $df(num)$ = _____

62. degrees of freedom – denominator: $df(denom)$ = _____

63. F statistic = _____

13.3 Facts About the F Distribution

DIRECTIONS

Use a solution sheet to conduct the following hypothesis tests. The solution sheet can be found in **Appendix E**.

64. Three students, Linda, Tuan, and Javier, are given five laboratory rats each for a nutritional experiment. Each rat's weight is recorded in grams. Linda feeds her rats Formula A, Tuan feeds his rats Formula B, and Javier feeds his rats Formula C. At the end of a specified time period, each rat is weighed again, and the net gain in grams is recorded. Using a significance level of 10%, test the hypothesis that the three formulas produce the same mean weight gain.

Linda's rats	Tuan's rats	Javier's rats
43.5	47.0	51.2
39.4	40.5	40.9
41.3	38.9	37.9
46.0	46.3	45.0
38.2	44.2	48.6

Table 13.21 Weights of Student Lab Rats

65. A grassroots group opposed to a proposed increase in the gas tax claimed that the increase would hurt working-class people the most, since they commute the farthest to work. Suppose that the group randomly surveyed 24 individuals and asked them their daily one-way commuting mileage. The results are in **Table 13.22**. Using a 5% significance level, test the hypothesis that the three mean commuting mileages are the same.

working-class	professional (middle incomes)	professional (wealthy)
17.8	16.5	8.5
26.7	17.4	6.3
49.4	22.0	4.6
9.4	7.4	12.6
65.4	9.4	11.0
47.1	2.1	28.6
19.5	6.4	15.4
51.2	13.9	9.3

Table 13.22

66. Examine the seven practice laps from **Table 13.1**. Determine whether the mean lap time is statistically the same for the seven practice laps, or if there is at least one lap that has a different mean time from the others.

Use the following information to answer the next two exercises. **Table 13.23** lists the number of pages in four different types of magazines.

home decorating	news	health	computer
172	87	82	104
286	94	153	136
163	123	87	98
205	106	103	207
197	101	96	146

Table 13.23

67. Using a significance level of 5%, test the hypothesis that the four magazine types have the same mean length.

68. Eliminate one magazine type that you now feel has a mean length different from the others. Redo the hypothesis test, testing that the remaining three means are statistically the same. Use a new solution sheet. Based on this test, are the mean lengths for the remaining three magazines statistically the same?

69. A researcher wants to know if the mean times (in minutes) that people watch their favorite news station are the same. Suppose that **Table 13.24** shows the results of a study.

CNN	FOX	Local
45	15	72
12	43	37
18	68	56
38	50	60
23	31	51
35	22	

Table 13.24

Assume that all distributions are normal, the four population standard deviations are approximately the same, and the data were collected independently and randomly. Use a level of significance of 0.05.

70. Are the means for the final exams the same for all statistics class delivery types? **Table 13.25** shows the scores on final exams from several randomly selected classes that used the different delivery types.

Online	Hybrid	Face-to-Face
72	83	80
84	73	78
77	84	84
80	81	81
81		86
		79
		82

Table 13.25

Assume that all distributions are normal, the four population standard deviations are approximately the same, and the data were collected independently and randomly. Use a level of significance of 0.05.

71. Are the mean number of times a month a person eats out the same for whites, blacks, Hispanics and Asians? Suppose that **Table 13.26** shows the results of a study.

White	Black	Hispanic	Asian
6	4	7	8
8	1	3	3
2	5	5	5
4	2	4	1
6		6	7

Table 13.26

Assume that all distributions are normal, the four population standard deviations are approximately the same, and the data were collected independently and randomly. Use a level of significance of 0.05.

72. Are the mean numbers of daily visitors to a ski resort the same for the three types of snow conditions? Suppose that **Table 13.27** shows the results of a study.

Powder	Machine Made	Hard Packed
1,210	2,107	2,846
1,080	1,149	1,638
1,537	862	2,019
941	1,870	1,178
	1,528	2,233
	1,382	

Table 13.27

Assume that all distributions are normal, the four population standard deviations are approximately the same, and the data were collected independently and randomly. Use a level of significance of 0.05.

73. Sanjay made identical paper airplanes out of three different weights of paper, light, medium and heavy. He made four airplanes from each of the weights, and launched them himself across the room. Here are the distances (in meters) that his planes flew.

Paper Type/Trial	Trial 1	Trial 2	Trial 3	Trial 4
Heavy	5.1 meters	3.1 meters	4.7 meters	5.3 meters
Medium	4 meters	3.5 meters	4.5 meters	6.1 meters
Light	3.1 meters	3.3 meters	2.1 meters	1.9 meters

Table 13.28

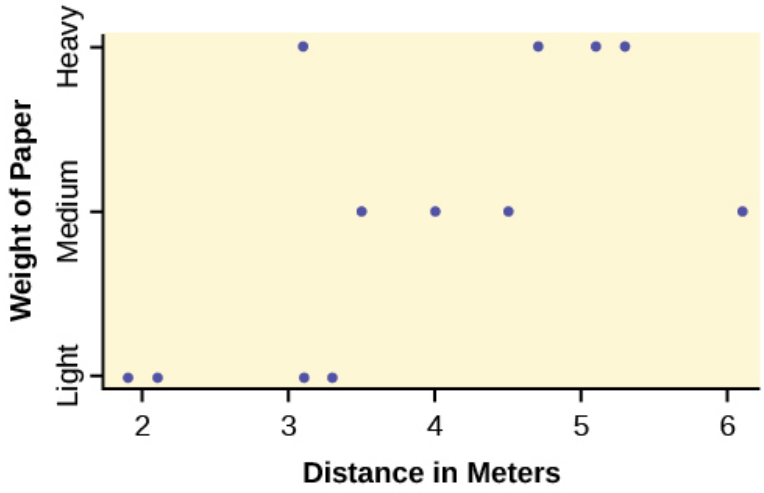

Figure 13.8

a. Take a look at the data in the graph. Look at the spread of data for each group (light, medium, heavy). Does it seem reasonable to assume a normal distribution with the same variance for each group? Yes or No.
b. Why is this a balanced design?
c. Calculate the sample mean and sample standard deviation for each group.
d. Does the weight of the paper have an effect on how far the plane will travel? Use a 1% level of significance. Complete the test using the method shown in the bean plant example in **Example 13.4**.
   - variance of the group means _____
   - $MS_{between}$= _____
   - mean of the three sample variances _____
   - MS_{within} = _____
   - F statistic = _____
   - $df(num)$ = _____, $df(denom)$ = _____
   - number of groups _____
   - number of observations _____
   - p-value = _____ ($P(F >$ _____) = _____)
   - Graph the p-value.
   - decision: _____
   - conclusion: _____

74. DDT is a pesticide that has been banned from use in the United States and most other areas of the world. It is quite effective, but persisted in the environment and over time became seen as harmful to higher-level organisms. Famously, egg shells of eagles and other raptors were believed to be thinner and prone to breakage in the nest because of ingestion of DDT in the food chain of the birds.

An experiment was conducted on the number of eggs (fecundity) laid by female fruit flies. There are three groups of flies. One group was bred to be resistant to DDT (the RS group). Another was bred to be especially susceptible to DDT (SS). Finally there was a control line of non-selected or typical fruitflies (NS). Here are the data:

RS	SS	NS	RS	SS	NS
12.8	38.4	35.4	22.4	23.1	22.6

Table 13.29

RS	SS	NS	RS	SS	NS
21.6	32.9	27.4	27.5	29.4	40.4
14.8	48.5	19.3	20.3	16	34.4
23.1	20.9	41.8	38.7	20.1	30.4
34.6	11.6	20.3	26.4	23.3	14.9
19.7	22.3	37.6	23.7	22.9	51.8
22.6	30.2	36.9	26.1	22.5	33.8
29.6	33.4	37.3	29.5	15.1	37.9
16.4	26.7	28.2	38.6	31	29.5
20.3	39	23.4	44.4	16.9	42.4
29.3	12.8	33.7	23.2	16.1	36.6
14.9	14.6	29.2	23.6	10.8	47.4
27.3	12.2	41.7			

Table 13.29

The values are the average number of eggs laid daily for each of 75 flies (25 in each group) over the first 14 days of their lives. Using a 1% level of significance, are the mean rates of egg selection for the three strains of fruitfly different? If so, in what way? Specifically, the researchers were interested in whether or not the selectively bred strains were different from the nonselected line, and whether the two selected lines were different from each other.

Here is a chart of the three groups:

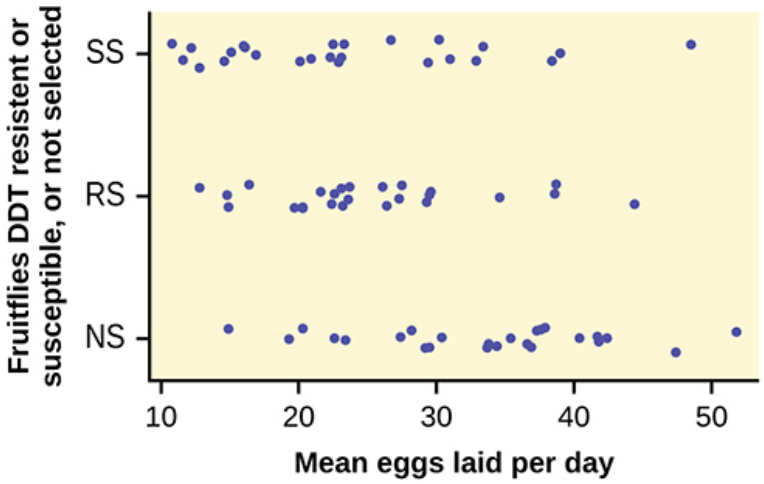

Figure 13.9

75. The data shown is the recorded body temperatures of 130 subjects as estimated from available histograms.

Traditionally we are taught that the normal human body temperature is 98.6 F. This is not quite correct for everyone. Are the mean temperatures among the four groups different?

Calculate 95% confidence intervals for the mean body temperature in each group and comment about the confidence intervals.

FL	FH	ML	MH	FL	FH	ML	MH
96.4	96.8	96.3	96.9	98.4	98.6	98.1	98.6
96.7	97.7	96.7	97	98.7	98.6	98.1	98.6

Table 13.30

FL	FH	ML	MH	FL	FH	ML	MH
97.2	97.8	97.1	97.1	98.7	98.6	98.2	98.7
97.2	97.9	97.2	97.1	98.7	98.7	98.2	98.8
97.4	98	97.3	97.4	98.7	98.7	98.2	98.8
97.6	98	97.4	97.5	98.8	98.8	98.2	98.8
97.7	98	97.4	97.6	98.8	98.8	98.3	98.9
97.8	98	97.4	97.7	98.8	98.8	98.4	99
97.8	98.1	97.5	97.8	98.8	98.9	98.4	99
97.9	98.3	97.6	97.9	99.2	99	98.5	99
97.9	98.3	97.6	98	99.3	99	98.5	99.2
98	98.3	97.8	98		99.1	98.6	99.5
98.2	98.4	97.8	98		99.1	98.6	
98.2	98.4	97.8	98.3		99.2	98.7	
98.2	98.4	97.9	98.4		99.4	99.1	
98.2	98.4	98	98.4		99.9	99.3	
98.2	98.5	98	98.6		100	99.4	
98.2	98.6	98	98.6		100.8		

Table 13.30

13.4 Test of Two Variances

76. Three students, Linda, Tuan, and Javier, are given five laboratory rats each for a nutritional experiment. Each rat's weight is recorded in grams. Linda feeds her rats Formula A, Tuan feeds his rats Formula B, and Javier feeds his rats Formula C. At the end of a specified time period, each rat is weighed again and the net gain in grams is recorded.

Linda's rats	Tuan's rats	Javier's rats
43.5	47.0	51.2
39.4	40.5	40.9
41.3	38.9	37.9
46.0	46.3	45.0
38.2	44.2	48.6

Table 13.31

Determine whether or not the variance in weight gain is statistically the same among Javier's and Linda's rats. Test at a significance level of 10%.

77. A grassroots group opposed to a proposed increase in the gas tax claimed that the increase would hurt working-class people the most, since they commute the farthest to work. Suppose that the group randomly surveyed 24 individuals and asked them their daily one-way commuting mileage. The results are as follows.

working-class	professional (middle incomes)	professional (wealthy)
17.8	16.5	8.5
26.7	17.4	6.3

Table 13.32

working-class	professional (middle incomes)	professional (wealthy)
49.4	22.0	4.6
9.4	7.4	12.6
65.4	9.4	11.0
47.1	2.1	28.6
19.5	6.4	15.4
51.2	13.9	9.3

Table 13.32

Determine whether or not the variance in mileage driven is statistically the same among the working class and professional (middle income) groups. Use a 5% significance level.

78. Refer to the data from Table 13.1.

Examine practice laps 3 and 4. Determine whether or not the variance in lap time is statistically the same for those practice laps.

Use the following information to answer the next two exercises. The following table lists the number of pages in four different types of magazines.

home decorating	news	health	computer
172	87	82	104
286	94	153	136
163	123	87	98
205	106	103	207
197	101	96	146

Table 13.33

79. Which two magazine types do you think have the same variance in length?

80. Which two magazine types do you think have different variances in length?

81. Is the variance for the amount of money, in dollars, that shoppers spend on Saturdays at the mall the same as the variance for the amount of money that shoppers spend on Sundays at the mall? Suppose that the **Table 13.34** shows the results of a study.

Saturday	Sunday	Saturday	Sunday
75	44	62	137
18	58	0	82
150	61	124	39
94	19	50	127
62	99	31	141
73	60	118	73
	89		

Table 13.34

82. Are the variances for incomes on the East Coast and the West Coast the same? Suppose that **Table 13.35** shows the results of a study. Income is shown in thousands of dollars. Assume that both distributions are normal. Use a level of significance of 0.05.

East	West
38	71
47	126
30	42
82	51
75	44
52	90
115	88
67	

Table 13.35

83. Thirty men in college were taught a method of finger tapping. They were randomly assigned to three groups of ten, with each receiving one of three doses of caffeine: 0 mg, 100 mg, 200 mg. This is approximately the amount in no, one, or two cups of coffee. Two hours after ingesting the caffeine, the men had the rate of finger tapping per minute recorded. The experiment was double blind, so neither the recorders nor the students knew which group they were in. Does caffeine affect the rate of tapping, and if so how?

Here are the data:

0 mg	100 mg	200 mg	0 mg	100 mg	200 mg
242	248	246	245	246	248
244	245	250	248	247	252
247	248	248	248	250	250
242	247	246	244	246	248
246	243	245	242	244	250

Table 13.36

84. King Manuel I, Komnenus ruled the Byzantine Empire from Constantinople (Istanbul) during the years 1145 to 1180 A.D. The empire was very powerful during his reign, but declined significantly afterwards. Coins minted during his era were found in Cyprus, an island in the eastern Mediterranean Sea. Nine coins were from his first coinage, seven from the second, four from the third, and seven from a fourth. These spanned most of his reign. We have data on the silver content of the coins:

First Coinage	Second Coinage	Third Coinage	Fourth Coinage
5.9	6.9	4.9	5.3
6.8	9.0	5.5	5.6
6.4	6.6	4.6	5.5
7.0	8.1	4.5	5.1
6.6	9.3		6.2
7.7	9.2		5.8
7.2	8.6		5.8
6.9			
6.2			

Table 13.37

Did the silver content of the coins change over the course of Manuel's reign?

Here are the means and variances of each coinage. The data are unbalanced.

	First	Second	Third	Fourth
Mean	6.7444	8.2429	4.875	5.6143
Variance	0.2953	1.2095	0.2025	0.1314

Table 13.38

85. The American League and the National League of Major League Baseball are each divided into three divisions: East, Central, and West. Many years, fans talk about some divisions being stronger (having better teams) than other divisions. This may have consequences for the postseason. For instance, in 2012 Tampa Bay won 90 games and did not play in the postseason, while Detroit won only 88 and did play in the postseason. This may have been an oddity, but is there good evidence that in the 2012 season, the American League divisions were significantly different in overall records? Use the following data to test whether the mean number of wins per team in the three American League divisions were the same or not. Note that the data are not balanced, as two divisions had five teams, while one had only four.

Division	Team	Wins
East	NY Yankees	95
East	Baltimore	93
East	Tampa Bay	90
East	Toronto	73
East	Boston	69

Table 13.39

Division	Team	Wins
Central	Detroit	88
Central	Chicago Sox	85
Central	Kansas City	72
Central	Cleveland	68
Central	Minnesota	66

Table 13.40

Division	Team	Wins
West	Oakland	94
West	Texas	93
West	LA Angels	89
West	Seattle	75

Table 13.41

REFERENCES

13.2 The F Distribution and the F-Ratio

Tomato Data, Marist College School of Science (unpublished student research)

13.3 Facts About the F Distribution

Data from a fourth grade classroom in 1994 in a private K – 12 school in San Jose, CA.

Hand, D.J., F. Daly, A.D. Lunn, K.J. McConway, and E. Ostrowski. *A Handbook of Small Datasets: Data for Fruitfly Fecundity.* London: Chapman & Hall, 1994.

Hand, D.J., F. Daly, A.D. Lunn, K.J. McConway, and E. Ostrowski. *A Handbook of Small Datasets.* London: Chapman & Hall, 1994, pg. 50.

Hand, D.J., F. Daly, A.D. Lunn, K.J. McConway, and E. Ostrowski. A Handbook of Small Datasets. London: Chapman & Hall, 1994, pg. 118.

"MLB Standings – 2012." Available online at http://espn.go.com/mlb/standings/_/year/2012.

Mackowiak, P. A., Wasserman, S. S., and Levine, M. M. (1992), "A Critical Appraisal of 98.6 Degrees F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlich," *Journal of the American Medical Association*, 268, 1578-1580.

13.4 Test of Two Variances

"MLB Vs. Division Standings – 2012." Available online at http://espn.go.com/mlb/standings/_/year/2012/type/vs-division/order/true.

SOLUTIONS

1 Each population from which a sample is taken is assumed to be normal.

3 The populations are assumed to have equal standard deviations (or variances).

5 The response is a numerical value.

7 H_a: At least two of the group means μ_1, μ_2, μ_3 are not equal.

9 4,939.2

11 2

13 2,469.6

15 3.7416

17 3

19 13.2

21 0.825

23 Because a one-way ANOVA test is always right-tailed, a high F statistic corresponds to a low p-value, so it is likely that we will reject the null hypothesis.

25 The curves approximate the normal distribution.

27 ten

29 $SS = 237.33$; $MS = 23.73$

31 0.1614

33 two

35 $SS = 5{,}700.4$; $MS = 2{,}850.2$

37 3.6101

39 Yes, there is enough evidence to show that the scores among the groups are statistically significant at the 10% level.

43 The populations from which the two samples are drawn are normally distributed.

45 H_0: $\sigma_1 = \sigma_2$ H_a: $\sigma_1 < \sigma_2$ or H_0: $\sigma_1^2 = \sigma_2^2$ H_a: $\sigma_1^2 < \sigma_2^2$

47 4.11

49 0.7159

51 No, at the 10% level of significance, we do not reject the null hypothesis and state that the data do not show that the variation in drive times for the first worker is less than the variation in drive times for the second worker.

53 2.8674

55 Reject the null hypothesis. There is enough evidence to say that the variance of the grades for the first student is higher than the variance in the grades for the second student.

57 0.7414

59 $SS_{between} = 26$
$SS_{within} = 441$
$F = 0.2653$

62 $df(denom) = 15$

64

a. H_0: $\mu_L = \mu_T = \mu_J$
b. at least any two of the means are different
c. $df(num) = 2$; $df(denom) = 12$
d. F distribution
e. 0.67
f. 0.5305
g. Check student's solution.
h. Decision: Do not reject null hypothesis; Conclusion: There is insufficient evidence to conclude that the means are different.

66

a. H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu_7$
b. At least two mean lap times are different.
c. $df(num) = 6$; $df(denom) = 98$
d. F distribution
e. 1.69
f. 0.1319
g. Check student's solution.
h. Decision: Do not reject null hypothesis; Conclusion: There is insufficient evidence to conclude that the mean lap times are different.

68

a. H_a: $\mu_d = \mu_n = \mu_h$
b. At least any two of the magazines have different mean lengths.
c. $df(num) = 2$, $df(denom) = 12$
d. F distribtuion
e. $F = 15.28$

f. p-value = 0.001

g. Check student's solution.

h.
   i. Alpha: 0.05
   ii. Decision: Reject the Null Hypothesis.
   iii. Reason for decision: p-value < alpha
   iv. Conclusion: There is sufficient evidence to conclude that the mean lengths of the magazines are different.

70

a. H_0: $\mu_o = \mu_h = \mu_f$

b. At least two of the means are different.

c. $df(n) = 2$, $df(d) = 13$

d. $F_{2,13}$

e. 0.64

f. 0.5437

g. Check student's solution.

h.
   i. Alpha: 0.05
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: p-value > alpha
   iv. Conclusion: The mean scores of different class delivery are not different.

72

a. H_0: $\mu_p = \mu_m = \mu_h$

b. At least any two of the means are different.

c. $df(n) = 2$, $df(d) = 12$

d. $F_{2,12}$

e. 3.13

f. 0.0807

g. Check student's solution.

h.
   i. Alpha: 0.05
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: p-value > alpha
   iv. Conclusion: There is not sufficient evidence to conclude that the mean numbers of daily visitors are different.

74 The data appear normally distributed from the chart and of similar spread. There do not appear to be any serious outliers, so we may proceed with our ANOVA calculations, to see if we have good evidence of a difference between the three groups. H_0: $\mu_1 = \mu_2 = \mu_3$; H_a: $\mu_i \neq \mu_j$ some $i \neq j$. Define μ_1, μ_2, μ_3, as the population mean number of eggs laid by the three groups of fruit flies. F statistic = 8.6657; p-value = 0.0004

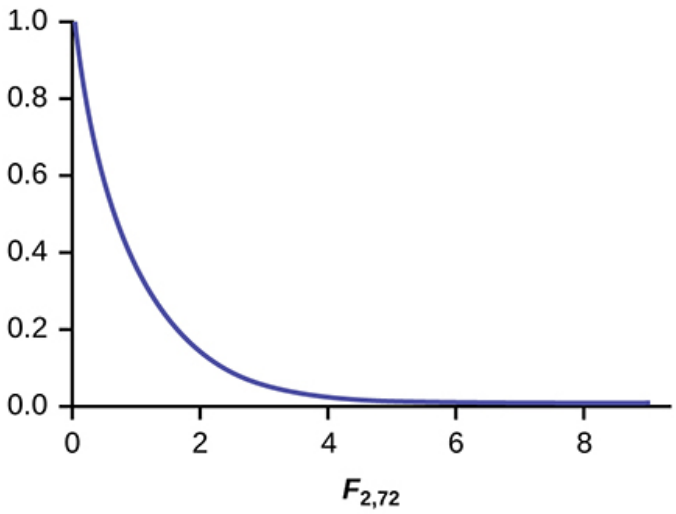

Figure 13.10

Decision: Since the *p*-value is less than the level of significance of 0.01, we reject the null hypothesis. **Conclusion:** We have good evidence that the average number of eggs laid during the first 14 days of life for these three strains of fruitflies are different. Interestingly, if you perform a two sample *t*-test to compare the RS and NS groups they are significantly different ($p = 0.0013$). Similarly, SS and NS are significantly different ($p = 0.0006$). However, the two selected groups, RS and SS are *not* significantly different ($p = 0.5176$). Thus we appear to have good evidence that selection either for resistance or for susceptibility involves a reduced rate of egg production (for these specific strains) as compared to flies that were not selected for resistance or susceptibility to DDT. Here, genetic selection has apparently involved a loss of fecundity.

76

a. $H_0: \sigma_1^2 = \sigma_2^2$

b. $H_a: \sigma_1^2 \neq \sigma_1^2$

c. $df(num) = 4$; $df(denom) = 4$

d. $F_{4,4}$

e. 3.00

f. 2(0.1563) = 0.3126. Using the TI-83+/84+ function 2-SampFtest, you get the test statistic as 2.9986 and *p*-value directly as 0.3127. If you input the lists in a different order, you get a test statistic of 0.3335 but the *p*-value is the same because this is a two-tailed test.

g. Check student's solution.

h. Decision: Do not reject the null hypothesis; Conclusion: There is insufficient evidence to conclude that the variances are different.

78

a. $H0: \sigma_1^2 = \sigma_2^2$

b. $Ha: \sigma_1^2 \neq \sigma_1^2$

c. $df(n) = 19$, $df(d) = 19$

d. $F_{19,19}$

e. 1.13

f. 0.786

g. Check student's solution.

h. i. Alpha:0.05

ii. Decision: Do not reject the null hypothesis.
iii. Reason for decision: *p*-value > alpha
iv. Conclusion: There is not sufficient evidence to conclude that the variances are different.

80 The answers may vary. Sample answer: Home decorating magazines and news magazines have different variances.

82

a. H_0: = $\sigma_1^2 = \sigma_2^2$
b. H_a: $\sigma_1^2 \neq \sigma_1^2$
c. $df(n) = 7$, $df(d) = 6$
d. $F_{7,6}$
e. 0.8117
f. 0.7825
g. Check student's solution.
h. i. Alpha: 0.05
ii. Decision: Do not reject the null hypothesis.
iii. Reason for decision: *p*-value > alpha
iv. Conclusion: There is not sufficient evidence to conclude that the variances are different.

84 Here is a strip chart of the silver content of the coins:

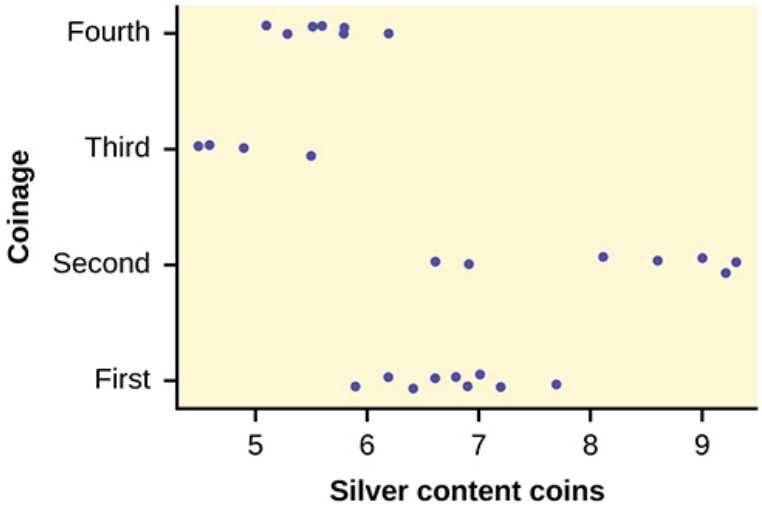

Figure 13.11

While there are differences in spread, it is not unreasonable to use ANOVA techniques. Here is the completed ANOVA table:

Source of Variation	Sum of Squares (*SS*)	Degrees of Freedom (*df*)	Mean Square (*MS*)	*F*
Factor (Between)	37.748	4 – 1 = 3	12.5825	26.272
Error (Within)	11.015	27 – 4 = 23	0.4789	
Total	48.763	27 – 1 = 26		

Table 13.42

$P(F > 26.272) = 0$; Reject the null hypothesis for any alpha. There is sufficient evidence to conclude that the mean silver content among the four coinages are different. From the strip chart, it appears that the first and second coinages had higher silver contents than the third and fourth.

85 Here is a stripchart of the number of wins for the 14 teams in the AL for the 2012 season.

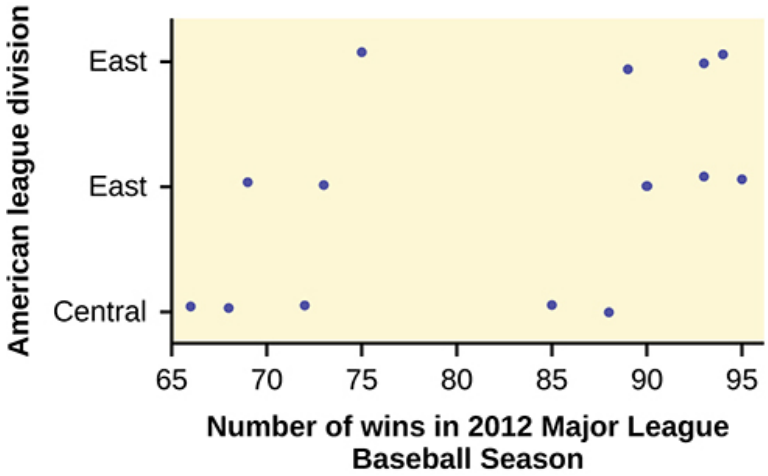

Figure 13.12

While the spread seems similar, there may be some question about the normality of the data, given the wide gaps in the middle near the 0.500 mark of 82 games (teams play 162 games each season in MLB). However, one-way ANOVA is robust. Here is the ANOVA table for the data:

Source of Variation	Sum of Squares (*SS*)	Degrees of Freedom (*df*)	Mean Square (*MS*)	*F*
Factor (Between)	344.16	3 – 1 = 2	172.08	26.272
Error (Within)	1,219.55	14 – 3 = 11	110.87	1.5521
Total	1,563.71	14 – 1 = 13		

Table 13.43

$P(F > 1.5521) = 0.2548$
Since the *p*-value is so large, there is not good evidence against the null hypothesis of equal means. We decline to reject the null hypothesis. Thus, for 2012, there is not any have any good evidence of a significant difference in mean number of wins between the divisions of the American League.

APPENDIX A: REVIEW EXERCISES (CH 3-13)

These review exercises are designed to provide extra practice on concepts learned before a particular chapter. For example, the review exercises for Chapter 3, cover material learned in chapters 1 and 2.

Chapter 3

Use the following information to answer the next six exercises: In a survey of 100 stocks on NASDAQ, the average percent increase for the past year was 9% for NASDAQ stocks.

1. The "average increase" for all NASDAQ stocks is the:

a. population
b. statistic
c. parameter
d. sample
e. variable

2. All of the NASDAQ stocks are the:

a. population
b. statistics
c. parameter
d. sample
e. variable

3. Nine percent is the:

a. population
b. statistics
c. parameter
d. sample
e. variable

4. The 100 NASDAQ stocks in the survey are the:

a. population
b. statistic
c. parameter
d. sample
e. variable

5. The percent increase for one stock in the survey is the:

a. population
b. statistic

c. parameter
d. sample
e. variable

6. Would the data collected by qualitative, quantitative discrete, or quantitative continuous?

Use the following information to answer the next two exercises: Thirty people spent two weeks around Mardi Gras in New Orleans. Their two-week weight gain is below. (Note: a loss is shown by a negative weight gain.)

Weight Gain	Frequency
–2	3
–1	5
0	2
1	4
4	13
6	2
11	1

Table A1

7. Calculate the following values:

a. the average weight gain for the two weeks
b. the standard deviation
c. the first, second, and third quartiles

8. Construct a histogram and box plot of the data.

Chapter 4

Use the following information to answer the next two exercises: A recent poll concerning credit cards found that 35 percent of respondents use a credit card that gives them a mile of air travel for every dollar they charge. Thirty percent of the respondents charge more than $2,000 per month. Of those respondents who charge more than $2,000, 80 percent use a credit card that gives them a mile of air travel for every dollar they charge.

9. What is the probability that a randomly selected respondent will spend more than $2,000 AND use a credit card that gives them a mile of air travel for every dollar they charge?

a. (0.30)(0.35)
b. (0.80)(0.35)
c. (0.80)(0.30)
d. (0.80)

10. Are using a credit card that gives a mile of air travel for each dollar spent AND charging more than $2,000 per month independent events?

a. Yes
b. No, and they are not mutually exclusive either.
c. No, but they are mutually exclusive.
d. Not enough information given to determine the answer

11. A sociologist wants to know the opinions of employed adult women about government funding for day care. She obtains a list of 520 members of a local business and professional women's club and mails a questionnaire to 100 of these women selected at random. Sixty-eight questionnaires are returned. What is the population in this study?

a. all employed adult women
b. all the members of a local business and professional women's club
c. the 100 women who received the questionnaire
d. all employed women with children

Use the following information to answer the next two exercises: The next two questions refer to the following: An article from The San Jose Mercury News was concerned with the racial mix of the 1500 students at Prospect High School in Saratoga, CA. The table summarizes the results. (Male and female values are approximate.) Suppose one Prospect High School student is randomly selected.

Gender/Ethnic group	White	Asian	Hispanic	Black	American Indian
Male	400	468	115	35	16
Female	440	132	140	40	14

Table A2

12. Find the probability that a student is Asian or Male.

13. Find the probability that a student is Black given that the student is female.

14. A sample of pounds lost, in a certain month, by individual members of a weight reducing clinic produced the following statistics:

- Mean = 5 lbs.
- Median = 4.5 lbs.
- Mode = 4 lbs.
- Standard deviation = 3.8 lbs.
- First quartile = 2 lbs.
- Third quartile = 8.5 lbs.

The correct statement is:

a. One fourth of the members lost exactly two pounds.
b. The middle fifty percent of the members lost from two to 8.5 lbs.
c. Most people lost 3.5 to 4.5 lbs.
d. All of the choices above are correct.

15. What does it mean when a data set has a standard deviation equal to zero?

a. All values of the data appear with the same frequency.
b. The mean of the data is also zero.
c. All of the data have the same value.
d. There are no data to begin with.

16. The statement that describe the illustration is:

Figure A1

a. the mean is equal to the median.
b. There is no first quartile.
c. The lowest data value is the median.
d. The median equals $\frac{Q_1 + Q_3}{2}$.

17. According to a recent article in the *San Jose Mercury News* the average number of babies born with significant hearing loss (deafness) is approximately 2 per 1000 babies in a healthy baby nursery. The number climbs to an average of 30 per 1000 babies in an intensive care nursery. Suppose that 1,000 babies from healthy baby nurseries were randomly surveyed. Find the probability that exactly two babies were born deaf.

18. A "friend" offers you the following "deal." For a $10 fee, you may pick an envelope from a box containing 100 seemingly identical envelopes. However, each envelope contains a coupon for a free gift.

- Ten of the coupons are for a free gift worth $6.
- Eighty of the coupons are for a free gift worth $8.
- Six of the coupons are for a free gift worth $12.
- Four of the coupons are for a free gift worth $40.

Based upon the financial gain or loss over the long run, should you play the game?

a. Yes, I expect to come out ahead in money.
b. No, I expect to come out behind in money.
c. It doesn't matter. I expect to break even.

Use the following information to answer the next four exercises: Recently, a nurse commented that when a patient calls the medical advice line claiming to have the flu, the chance that he/she truly has the flu (and not just a nasty cold) is only about 4%. Of the next 25 patients calling in claiming to have the flu, we are interested in how many actually have the flu.

19. Define the random variable and list its possible values.

20. State the distribution of X.

21. Find the probability that at least four of the 25 patients actually have the flu.

22. On average, for every 25 patients calling in, how many do you expect to have the flu?

Use the following information to answer the next two exercises: Different types of writing can sometimes be distinguished by the number of letters in the words used. A student interested in this fact wants to study the number of letters of words used by Tom Clancy in his novels. She opens a Clancy novel at random and records the number of letters of the first 250 words on the page.

23. What kind of data was collected?

a. qualitative
b. quantitative continuous
c. quantitative discrete

24. What is the population under study?

Chapter 5

Use the following information to answer the next seven exercises: A recent study of mothers of junior high school children in Santa Clara County reported that 76% of the mothers are employed in paid positions. Of those mothers who are employed, 64% work full-time (over 35 hours per week), and 36% work part-time. However, out of all of the mothers in the population, 49% work full-time. The population under study is made up of mothers of junior high school children in Santa Clara County. Let E = employed and F = full-time employment.

25.

a. Find the percent of all mothers in the population that are NOT employed.

b. Find the percent of mothers in the population that are employed part-time.

26. The "type of employment" is considered to be what type of data?

27. Find the probability that a randomly selected mother works part-time given that she is employed.

28. Find the probability that a randomly selected person from the population will be employed or work full-time.

29. Being employed and working part-time:

a. mutually exclusive events? Why or why not?

b. independent events? Why or why not?

Use the following additional information to answer the next two exercises: We randomly pick ten mothers from the above population. We are interested in the number of the mothers that are employed. Let X = number of mothers that are employed.

30. State the distribution for X.

31. Find the probability that at least six are employed.

32. We expect the statistics discussion board to have, on average, 14 questions posted to it per week. We are interested in the number of questions posted to it per day.

a. Define X.

b. What are the values that the random variable may take on?

c. State the distribution for X.

d. Find the probability that from ten to 14 (inclusive) questions are posted to the listserv on a randomly picked day.

33. A person invests $1,000 into stock of a company that hopes to go public in one year. The probability that the person will lose all his money after one year (i.e. his stock will be worthless) is 35%. The probability that the person's stock will still have a value of $1,000 after one year (i.e. no profit and no loss) is 60%. The probability that the person's stock will increase in value by $10,000 after one year (i.e. will be worth $11,000) is 5%. Find the expected profit after one year.

34. Rachel's piano cost $3,000. The average cost for a piano is $4,000 with a standard deviation of $2,500. Becca's guitar cost $550. The average cost for a guitar is $500 with a standard deviation of $200. Matt's drums cost $600. The average cost for drums is $700 with a standard deviation of $100. Whose cost was lowest when compared to his or her own instrument?

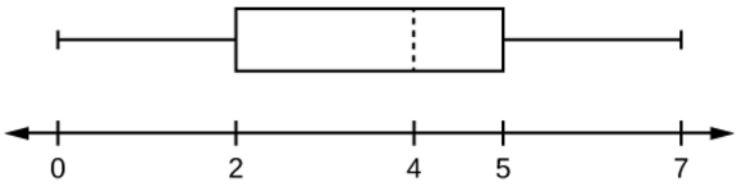

Figure A2

35. Explain why each statement is either true or false given the box plot in **Figure A2**.

a. Twenty-five percent of the data re at most five.

b. There is the same amount of data from 4–5 as there is from 5–7.

c. There are no data values of three.

d. Fifty percent of the data are four.

Using the following information to answer the next two exercises: 64 faculty members were asked the number of cars they owned (including spouse and children's cars). The results are given in the following graph:

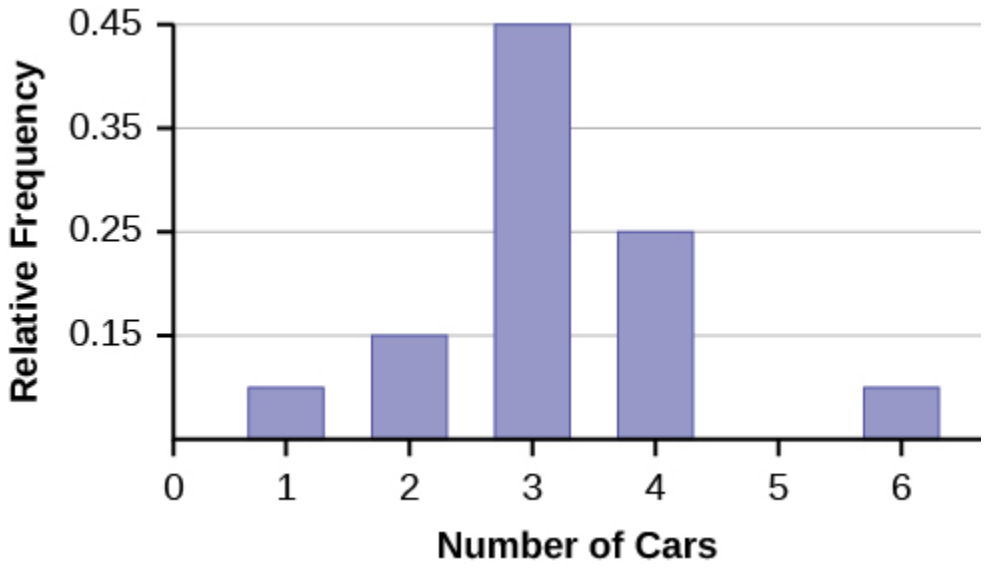

Figure A3

36. Find the approximate number of responses that were three.

37. Find the first, second and third quartiles. Use them to construct a box plot of the data.

Use the following information to answer the next three exercises: **Table A3** shows data gathered from 15 girls on the Snow Leopard soccer team when they were asked how they liked to wear their hair. Supposed one girl from the team is randomly selected.

Hair Style/Hair Color	Blond	Brown	Black
Ponytail	3	2	5
Plain	2	2	1

Table A3

38. Find the probability that the girl has black hair GIVEN that she wears a ponytail.

39. Find the probability that the girl wears her hair plain OR has brown hair.

40. Find the probability that the girl has blond hair AND that she wears her hair plain.

Chapter 6

Use the following information to answer the next two exercises: $X \sim U(3, 13)$

41. Explain which of the following are false and which are true.

a. $f(x) = \frac{1}{10}, 3 \leq x \leq 13$

b. There is no mode

c. The median is less than the mean.

d. $P(x > 10) = P(x \leq 6)$

42. Calculate:

a. the mean.

b. the median.

c. the 65^{th} percentile.

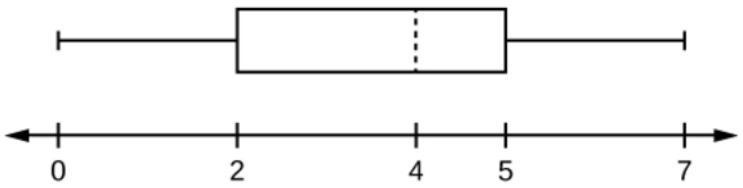

Figure A4

43. Which of the following is true for the box plot in **Figure A4**?

a. Twenty-five percent of the data are at most five.

b. There is about the same amount of data from 4–5 as there is from 5–7.

c. There are no data values of three.

d. Fifty percent of the data are four.

44. If $P(G|H) = P(G)$, then which of the following is correct?

a. G and H are mutually exclusive events.

b. $P(G) = P(H)$

c. Knowing that H has occurred will affect the chance that G will happen.

d. G and H are independent events.

45. If $P(J) = 0.3$, $P(K) = 0.63$, and J and K are independent events, then explain which are correct and which are incorrect.

a. $P(J \text{ AND } K) = 0$

b. $P(J \text{ OR } K) = 0.9$

c. $P(J \text{ OR } K) = 0.72$

d. $P(J) \neq P(J|K)$

46. On average, five students from each high school class get full scholarships to four-year colleges. Assume that most high school classes have about 500 students. X = the number of students from a high school class that get full scholarships to four-year schools. Which of the following is the distribution of X?

a. $P(5)$

b. $B(500, 5)$

c. $Exp\left(\frac{1}{5}\right)$

d. $N\left(5, \frac{(0.01)(0.99)}{500}\right)$

Chapter 7

Use the following information to answer the next three exercises: Richard's Furniture Company delivers furniture from 10 A.M. to 2 P.M. continuously and uniformly. We are interested in how long (in hours) past the 10 A.M. start time that individuals wait for their delivery.

47. $X \sim$ _____

a. $U(0, 4)$

b. $U(10, 20)$
c. $Exp(2)$
d. $N(2, 1)$

48. The average wait time is:

a. 1 hour.
b. 2 hours.
c. 2.5 hours.
d. 4 hours.

49. Suppose that it is now past noon on a delivery day. The probability that a person must wait at least 1.5 more hours is:

a. $\frac{1}{4}$
b. $\frac{1}{2}$
c. $\frac{3}{4}$
d. $\frac{3}{8}$

50. Given: $X \sim Exp\left(\frac{1}{3}\right)$

a. Find $P(x > 1)$.
b. Calculate the minimum value for the upper quartile.
c. Find $P\left(x = \frac{1}{3}\right)$

51.

- 40% of full-time students took 4 years to graduate
- 30% of full-time students took 5 years to graduate
- 20% of full-time students took 6 years to graduate
- 10% of full-time students took 7 years to graduate

The expected time for full-time students to graduate is:

a. 4 years
b. 4.5 years
c. 5 years
d. 5.5 years

52. Which of the following distributions is described by the following example?
Many people can run a short distance of under two miles, but as the distance increases, fewer people can run that far.

a. binomial
b. uniform
c. exponential
d. normal

53. The length of time to brush one's teeth is generally thought to be exponentially distributed with a mean of $\frac{3}{4}$ minutes. Find the probability that a randomly selected person brushes his or her teeth less than $\frac{3}{4}$ minutes.

a. 0.5
b. $\frac{3}{4}$
c. 0.43
d. 0.63

54. Which distribution accurately describes the following situation?
The chance that a teenage boy regularly gives his mother a kiss goodnight is about 20%. Fourteen teenage boys are randomly surveyed. Let X = the number of teenage boys that regularly give their mother a kiss goodnight.

a. $B(14,0.20)$
b. $P(2.8)$
c. $N(2.8,2.24)$
d. $Exp\left(\frac{1}{0.20}\right)$

55. A 2008 report on technology use states that approximately 20% of U.S. households have never sent an e-mail. Suppose that we select a random sample of fourteen U.S. households. Let X = the number of households in a 2008 sample of 14 households that have never sent an email

a. $B(14,0.20)$
b. $P(2.8)$
c. $N(2.8,2.24)$
d. $Exp\left(\frac{1}{0.20}\right)$

Chapter 8

Use the following information to answer the next three exercises: Suppose that a sample of 15 randomly chosen people were put on a special weight loss diet. The amount of weight lost, in pounds, follows an unknown distribution with mean equal to 12 pounds and standard deviation equal to three pounds. Assume that the distribution for the weight loss is normal.

56. To find the probability that the mean amount of weight lost by 15 people is no more than 14 pounds, the random variable should be:

a. number of people who lost weight on the special weight loss diet.
b. the number of people who were on the diet.
c. the mean amount of weight lost by 15 people on the special weight loss diet.
d. the total amount of weight lost by 15 people on the special weight loss diet.

57. Find the probability asked for in **Question 56**.

58. Find the 90^{th} percentile for the mean amount of weight lost by 15 people.

Using the following information to answer the next three exercises: The time of occurrence of the first accident during rush-hour traffic at a major intersection is uniformly distributed between the three hour interval 4 p.m. to 7 p.m. Let X = the amount of time (hours) it takes for the first accident to occur.

59. What is the probability that the time of occurrence is within the first half-hour or the last hour of the period from 4 to 7 p.m.?

a. cannot be determined from the information given

b. $\frac{1}{6}$

c. $\frac{1}{2}$

d. $\frac{1}{3}$

60. The 20^{th} percentile occurs after how many hours?

a. 0.20

b. 0.60

c. 0.50

d. 1

61. Assume Ramon has kept track of the times for the first accidents to occur for 40 different days. Let C = the total cumulative time. Then C follows which distribution?

a. $U(0,3)$

b. $Exp(13)$

c. $N(60, 5.477)$

d. $N(1.5, 0.01875)$

62. Using the information in **Question 61**, find the probability that the total time for all first accidents to occur is more than 43 hours.

Use the following information to answer the next two exercises: The length of time a parent must wait for his children to clean their rooms is uniformly distributed in the time interval from one to 15 days.

63. How long must a parent expect to wait for his children to clean their rooms?

a. eight days

b. three days

c. 14 days

d. six days

64. What is the probability that a parent will wait more than six days given that the parent has already waited more than three days?

a. 0.5174

b. 0.0174

c. 0.7500

d. 0.2143

Use the following information to answer the next five exercises: Twenty percent of the students at a local community college live in within five miles of the campus. Thirty percent of the students at the same community college receive some kind of financial aid. Of those who live within five miles of the campus, 75% receive some kind of financial aid.

65. Find the probability that a randomly chosen student at the local community college does not live within five miles of the campus.

a. 80%

b. 20%

c. 30%

d. cannot be determined

66. Find the probability that a randomly chosen student at the local community college lives within five miles of the campus or receives some kind of financial aid.

a. 50%
b. 35%
c. 27.5%
d. 75%

67. Are living in student housing within five miles of the campus and receiving some kind of financial aid mutually exclusive?

a. yes
b. no
c. cannot be determined

68. The interest rate charged on the financial aid is _____ data.

a. quantitative discrete
b. quantitative continuous
c. qualitative discrete
d. qualitative

69. The following information is about the students who receive financial aid at the local community college.

- 1st quartile = $250
- 2nd quartile = $700
- 3rd quartile = $1200

These amounts are for the school year. If a sample of 200 students is taken, how many are expected to receive $250 or more?

a. 50
b. 250
c. 150
d. cannot be determined

Use the following information to answer the next two exercises: $P(A) = 0.2$, $P(B) = 0.3$; A and B are independent events.

70. $P(A \text{ AND } B)$ = _____

a. 0.5
b. 0.6
c. 0
d. 0.06

71. $P(A \text{ OR } B)$ = _____

a. 0.56
b. 0.5
c. 0.44
d. 1

72. If H and D are mutually exclusive events, $P(H) = 0.25$, $P(D) = 0.15$, then $P(H|D)$.

a. 1

b. 0

c. 0.40

d. 0.0375

Chapter 9

73. Rebecca and Matt are 14 year old twins. Matt's height is two standard deviations below the mean for 14 year old boys' height. Rebecca's height is 0.10 standard deviations above the mean for 14 year old girls' height. Interpret this.

a. Matt is 2.1 inches shorter than Rebecca.

b. Rebecca is very tall compared to other 14 year old girls.

c. Rebecca is taller than Matt.

d. Matt is shorter than the average 14 year old boy.

74. Construct a histogram of the IPO data (see **Appendix C**).

Use the following information to answer the next three exercises: Ninety homeowners were asked the number of estimates they obtained before having their homes fumigated. Let X = the number of estimates.

x	Relative Frequency	Cumulative Relative Frequency
1	0.3	
2	0.2	
4	0.4	
5	0.1	

Table A4

75. Complete the cumulative frequency column.

76. Calculate the sample mean (a), the sample standard deviation (b) and the percent of the estimates that fall at or below four (c).

77. Calculate the median, M, the first quartile, Q_1, the third quartile, Q_3. Then construct a box plot of the data.

78. The middle 50% of the data are between _____ and _____.

Use the following information to answer the next three exercises: Seventy 5^{th} and 6^{th} graders were asked their favorite dinner.

	Pizza	Hamburgers	Spaghetti	Fried shrimp
5th grader	15	6	9	0
6th grader	15	7	10	8

Table A5

79. Find the probability that one randomly chosen child is in the 6th grade and prefers fried shrimp.

a. $\frac{32}{70}$

b. $\frac{8}{32}$

c. $\frac{8}{8}$

d. $\frac{8}{70}$

80. Find the probability that a child does not prefer pizza.

a. $\frac{30}{70}$

b. $\frac{30}{40}$

c. $\frac{40}{70}$

d. 1

81. Find the probability a child is in the 5^{th} grade given that the child prefers spaghetti.

a. $\frac{9}{19}$

b. $\frac{9}{70}$

c. $\frac{9}{30}$

d. $\frac{19}{70}$

82. A sample of convenience is a random sample.

a. true

b. false

83. A statistic is a number that is a property of the population.

a. true

b. false

84. You should always throw out any data that are outliers.

a. true

b. false

85. Lee bakes pies for a small restaurant in Felton, CA. She generally bakes 20 pies in a day, on average. Of interest is the number of pies she bakes each day.

a. Define the random variable X.

b. State the distribution for X.

c. Find the probability that Lee bakes more than 25 pies in any given day.

86. Six different brands of Italian salad dressing were randomly selected at a supermarket. The grams of fat per serving are 7, 7, 9, 6, 8, 5. Assume that the underlying distribution is normal. Calculate a 95% confidence interval for the population mean grams of fat per serving of Italian salad dressing sold in supermarkets.

87. Given: uniform, exponential, normal distributions. Match each to a statement below.

a. mean = median ≠ mode

b. mean > median > mode

c. mean = median = mode

Chapter 10

Use the following information to answer the next three exercises: In a survey at Kirkwood Ski Resort the following information was recorded:

	0–10	11–20	21–40	40+
Ski	10	12	30	8
Snowboard	6	17	12	5

Table A6

Suppose that one person from **Table A6** was randomly selected.

88. Find the probability that the person was a skier or was age 11–20.

89. Find the probability that the person was a snowboarder given he or she was age 21–40.

90. Explain which of the following are true and which are false.

a. Sport and age are independent events.

b. Ski and age 11–20 are mutually exclusive events.

c. P(Ski AND age 21–40) < P(Ski|age 21–40)

d. P(Snowboard OR age 0–10) < P(Snowboard|age 0–10)

91. The average length of time a person with a broken leg wears a cast is approximately six weeks. The standard deviation is about three weeks. Thirty people who had recently healed from broken legs were interviewed. State the distribution that most accurately reflects total time to heal for the thirty people.

92. The distribution for X is uniform. What can we say for certain about the distribution for $\bar{X}$ when $n = 1$?

a. The distribution for $\bar{X}$ is still uniform with the same mean and standard deviation as the distribution for X.

b. The distribution for $\bar{X}$ is normal with the different mean and a different standard deviation as the distribution for X.

c. The distribution for $\bar{X}$ is normal with the same mean but a larger standard deviation than the distribution for X.

d. The distribution for $\bar{X}$ is normal with the same mean but a smaller standard deviation than the distribution for X.

93. The distribution for X is uniform. What can we say for certain about the distribution for $\sum X$ when $n = 50$?

a. distribution for $\sum X$ is still uniform with the same mean and standard deviation as the distribution for X.

b. The distribution for $\sum X$ is normal with the same mean but a larger standard deviation as the distribution for X.

c. The distribution for $\sum X$ is normal with a larger mean and a larger standard deviation than the distribution for X.

d. The distribution for $\sum X$ is normal with the same mean but a smaller standard deviation than the distribution for X.

Use the following information to answer the next three exercises: A group of students measured the lengths of all the carrots in a five-pound bag of baby carrots. They calculated the average length of baby carrots to be 2.0 inches with a standard deviation of 0.25 inches. Suppose we randomly survey 16 five-pound bags of baby carrots.

94. State the approximate distribution for $\bar{X}$, the distribution for the average lengths of baby carrots in 16 five-pound bags.

$\bar{X}$ ~ _____

95. Explain why we cannot find the probability that one individual randomly chosen carrot is greater than 2.25 inches.

96. Find the probability that $\bar{x}$ is between two and 2.25 inches.

Use the following information to answer the next three exercises: At the beginning of the term, the amount of time a student waits in line at the campus store is normally distributed with a mean of five minutes and a standard deviation of two minutes.

97. Find the 90^{th} percentile of waiting time in minutes.

98. Find the median waiting time for one student.

99. Find the probability that the average waiting time for 40 students is at least 4.5 minutes.

Chapter 11

Use the following information to answer the next four exercises: Suppose that the time that owners keep their cars (purchased new) is normally distributed with a mean of seven years and a standard deviation of two years. We are interested in how long an individual keeps his car (purchased new). Our population is people who buy their cars new.

100. Sixty percent of individuals keep their cars **at most** how many years?

101. Suppose that we randomly survey one person. Find the probability that person keeps his or her car **less than** 2.5 years.

102. If we are to pick individuals ten at a time, find the distribution for the **mean** car length ownership.

103. If we are to pick ten individuals, find the probability that the **sum** of their ownership time is more than 55 years.

104. For which distribution is the median not equal to the mean?

a. Uniform
b. Exponential
c. Normal
d. Student *t*

105. Compare the standard normal distribution to the Student's *t*-distribution, centered at zero. Explain which of the following are true and which are false.

a. As the number surveyed increases, the area to the left of –1 for the Student's *t*-distribution approaches the area for the standard normal distribution.
b. As the degrees of freedom decrease, the graph of the Student's *t*-distribution looks more like the graph of the standard normal distribution.
c. If the number surveyed is 15, the normal distribution should never be used.

Use the following information to answer the next five exercises: We are interested in the checking account balance of twenty-year-old college students. We randomly survey 16 twenty-year-old college students. We obtain a sample mean of \$640 and a sample standard deviation of \$150. Let X = checking account balance of an individual twenty year old college student.

106. Explain why we cannot determine the distribution of X.

107. If you were to create a confidence interval or perform a hypothesis test for the population mean checking account balance of twenty-year-old college students, what distribution would you use?

108. Find the 95% confidence interval for the true mean checking account balance of a twenty-year-old college student.

109. What type of data is the balance of the checking account considered to be?

110. What type of data is the number of twenty-year-olds considered to be?

111. On average, a busy emergency room gets a patient with a shotgun wound about once per week. We are interested in the number of patients with a shotgun wound the emergency room gets per 28 days.

a. Define the random variable X.
b. State the distribution for X.
c. Find the probability that the emergency room gets no patients with shotgun wounds in the next 28 days.

Use the following information to answer the next two exercises: The probability that a certain slot machine will pay back money when a quarter is inserted is 0.30. Assume that each play of the slot machine is independent from each other. A person puts in 15 quarters for 15 plays.

112. Is the expected number of plays of the slot machine that will pay back money greater than, less than or the same as the median? Explain your answer.

113. Is it likely that exactly eight of the 15 plays would pay back money? Justify your answer numerically.

114. A game is played with the following rules:

- it costs $10 to enter.
- a fair coin is tossed four times.
- if you do not get four heads or four tails, you lose your $10.
- if you get four heads or four tails, you get back your $10, plus $30 more.

Over the long run of playing this game, what are your expected earnings?

115.

- The mean grade on a math exam in Rachel's class was 74, with a standard deviation of five. Rachel earned an 80.
- The mean grade on a math exam in Becca's class was 47, with a standard deviation of two. Becca earned a 51.
- The mean grade on a math exam in Matt's class was 70, with a standard deviation of eight. Matt earned an 83.

Find whose score was the best, compared to his or her own class. Justify your answer numerically.

Use the following information to answer the next two exercises: A random sample of 70 compulsive gamblers were asked the number of days they go to casinos per week. The results are given in the following graph:

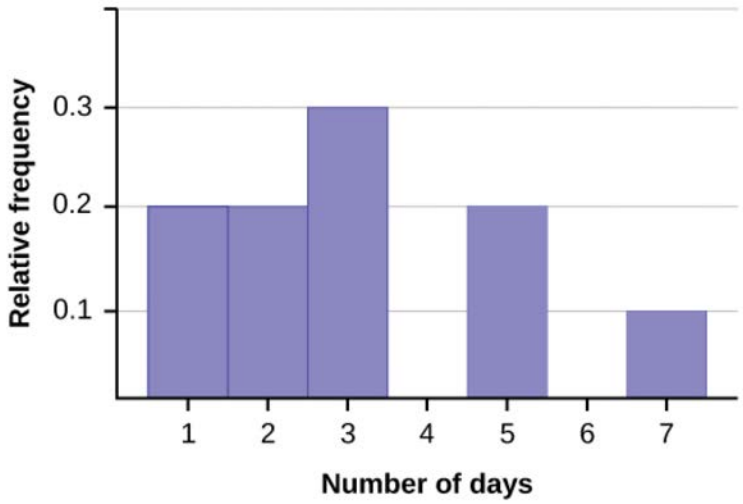

Figure A5

116. Find the number of responses that were five.

117. Find the mean, standard deviation, the median, the first quartile, the third quartile and the *IQR*.

118. Based upon research at De Anza College, it is believed that about 19% of the student population speaks a language other than English at home. Suppose that a study was done this year to see if that percent has decreased. Ninety-eight students were randomly surveyed with the following results. Fourteen said that they speak a language other than English at home.

a. State an appropriate null hypothesis.
b. State an appropriate alternative hypothesis.
c. Define the random variable, P'.
d. Calculate the test statistic.
e. Calculate the p-value.
f. At the 5% level of decision, what is your decision about the null hypothesis?
g. What is the Type I error?
h. What is the Type II error?

119. Assume that you are an emergency paramedic called in to rescue victims of an accident. You need to help a patient who is bleeding profusely. The patient is also considered to be a high risk for contracting AIDS. Assume that the null hypothesis is that the patient does **not** have the HIV virus. What is a Type I error?

120. It is often said that Californians are more casual than the rest of Americans. Suppose that a survey was done to see if the proportion of Californian professionals that wear jeans to work is greater than the proportion of non-Californian professionals. Fifty of each was surveyed with the following results. Fifteen Californians wear jeans to work and six non-Californians wear jeans to work.
Let C = Californian professional; NC = non-Californian professional

a. State appropriate null and alternate hypotheses.
b. Define the random variable.
c. Calculate the test statistic and p-value.
d. At the 5% significance level, what is your decision?
e. What is the Type I error?
f. What is the Type II error?

Use the following information to answer the next two exercises: A group of Statistics students have developed a technique that they feel will lower their anxiety level on statistics exams. They measured their anxiety level at the start of the quarter and again at the end of the quarter. Recorded is the paired data in that order: (1000, 900); (1200, 1050); (600, 700); (1300, 1100); (1000, 900); (900, 900).

121. This is a test of (pick the best answer):

a. large samples, independent means
b. small samples, independent means
c. dependent means

122. State the distribution to use for the test.

Chapter 12

Use the following information to answer the next two exercises: A recent survey of U.S. teenage pregnancy was answered by 720 girls, age 12–19. Six percent of the girls surveyed said they have been pregnant. We are interested in the true proportion of U.S. girls, age 12–19, who have been pregnant.

123. Find the 95% confidence interval for the true proportion of U.S. girls, age 12–19, who have been pregnant.

124. The report also stated that the results of the survey are accurate to within ±3.7% at the 95% confidence level. Suppose that a new study is to be done. It is desired to be accurate to within 2% of the 95% confidence level. What is the minimum number that should be surveyed?

125. Given: $X \sim Exp\left(\frac{1}{3}\right)$. Sketch the graph that depicts: $P(x > 1)$.

Use the following information to answer the next three exercises: The amount of money a customer spends in one trip to the supermarket is known to have an exponential distribution. Suppose the mean amount of money a customer spends in one trip to the supermarket is $72.

126. Find the probability that one customer spends less than $72 in one trip to the supermarket?

127. Suppose five customers pool their money. How much money altogether would you expect the five customers to spend in one trip to the supermarket (in dollars)?

128. State the distribution to use if you want to find the probability that the **mean** amount spent by five customers in one trip to the supermarket is less than $60.

Chapter 13

Use the following information to answer the next two exercises: Suppose that the probability of a drought in any independent year is 20%. Out of those years in which a drought occurs, the probability of water rationing is 10%. However, in any year, the probability of water rationing is 5%.

129. What is the probability of both a drought **and** water rationing occurring?

130. Out of the years with water rationing, find the probability that there is a drought.

Use the following information to answer the next three exercises:

	Apple	Pumpkin	Pecan
Female	40	10	30
Male	20	30	10

Table A7

131. Suppose that one individual is randomly chosen. Find the probability that the person's favorite pie is apple **or** the person is male.

132. Suppose that one male is randomly chosen. Find the probability his favorite pie is pecan.

133. Conduct a hypothesis test to determine if favorite pie type and gender are independent.

Use the following information to answer the next two exercises: Let's say that the probability that an adult watches the news at least once per week is 0.60.

134. We randomly survey 14 people. On average, how many people do we expect to watch the news at least once per week?

135. We randomly survey 14 people. Of interest is the number that watch the news at least once per week. State the distribution of X. $X \sim$ _____

136. The following histogram is most likely to be a result of sampling from which distribution?

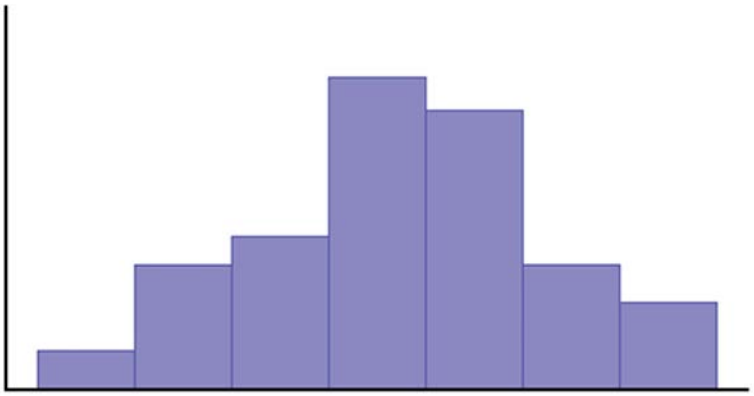

Figure A6

a. Chi-Square
b. Geometric
c. Uniform
d. Binomial

137. The ages of De Anza evening students is known to be normally distributed with a population mean of 40 and a population standard deviation of six. A sample of six De Anza evening students reported their ages (in years) as: 28; 35; 47; 45; 30; 50. Find the probability that the mean of six ages of randomly chosen students is less than 35 years. Hint: Find the sample mean.

138. A math exam was given to all the fifth grade children attending Country School. Two random samples of scores were taken. The null hypothesis is that the mean math scores for boys and girls in fifth grade are the same. Conduct a hypothesis test.

	n	$\bar{x}$	s^2
Boys	55	82	29
Girls	60	86	46

Table A8

139. In a survey of 80 males, 55 had played an organized sport growing up. Of the 70 females surveyed, 25 had played an organized sport growing up. We are interested in whether the proportion for males is higher than the proportion for females. Conduct a hypothesis test.

140. Which of the following is preferable when designing a hypothesis test?

a. Maximize α and minimize β
b. Minimize α and maximize β
c. Maximize α and β
d. Minimize α and β

Use the following information to answer the next three exercises: 120 people were surveyed as to their favorite beverage (non-alcoholic). The results are below.

Beverage/Age	0–9	10–19	20–29	30+	Totals
Milk	14	10	6	0	30
Soda	3	8	26	15	52
Juice	7	12	12	7	38
Totals	24	330	44	22	120

Table A9

141. Are the events of milk and 30+:

a. independent events? Justify your answer.
b. mutually exclusive events? Justify your answer.

142. Suppose that one person is randomly chosen. Find the probability that person is 10–19 given that he or she prefers juice.

143. Are "Preferred Beverage" and "Age" independent events? Conduct a hypothesis test.

144. Given the following histogram, which distribution is the data most likely to come from?

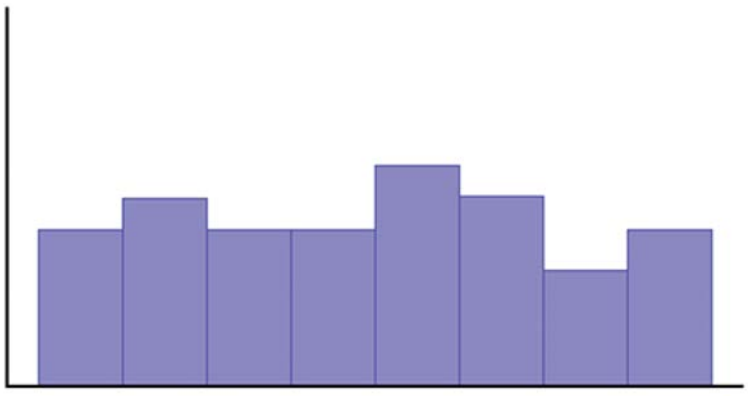

Figure A7

a. uniform
b. exponential
c. normal
d. chi-square

Solutions

Chapter 3

1. c. parameter

2. a. population

3. b. statistic

4. d. sample

5. e. variable

6. quantitative continuous

7.

a. 2.27
b. 3.04
c. –1, 4, 4

8. Answers will vary.

Chapter 4

9. c. (0.80)(0.30)

10. b. No, and they are not mutually exclusive either.

11. a. all employed adult women

12. 0.5773

13. 0.0522

14. b. The middle fifty percent of the members lost from 2 to 8.5 lbs.

15. c. All of the data have the same value.

16. c. The lowest data value is the median.

17. 0.279

18. b. No, I expect to come out behind in money.

19. X = the number of patients calling in claiming to have the flu, who actually have the flu.
X = 0, 1, 2, …25

20. $B(25, 0.04)$

21. 0.0165

22. 1

23. c. quantitative discrete

24. all words used by Tom Clancy in his novels

Chapter 5

25.

a. 24%
b. 27%

26. qualitative

27. 0.36

28. 0.7636

29.

a. No
b. No

30. $B(10, 0.76)$

31. 0.9330

32.

a. X = the number of questions posted to the statistics listserv per day.
b. $X = 0, 1, 2,\ldots$
c. $X \sim P(2)$
d. 0

33. $150

34. Matt

35.

a. false
b. true
c. false
d. false

36. 16

37. first quartile: 2
second quartile: 2
third quartile: 3

38. 0.5

39. $\frac{7}{15}$

40. $\frac{2}{15}$

Chapter 6

41.

a. true
b. true
c. False – the median and the mean are the same for this symmetric distribution.
d. true

42.

a. 8
b. 8
c. $P(x < k) = 0.65 = (k - 3)\left(\frac{1}{10}\right)$. $k = 9.5$

43.

a. False – $\frac{3}{4}$ of the data are at most five.

b. True – each quartile has 25% of the data.

c. False – that is unknown.

d. False – 50% of the data are four or less.

44. d. G and H are independent events.

45.

a. False – J and K are independent so they are not mutually exclusive which would imply dependency (meaning $P(J \text{ AND } K)$ is not 0).

b. False – see answer c.

c. True – $P(J \text{ OR } K) = P(J) + P(K) – P(J \text{ AND } K) = P(J) + P(K) – P(J)P(K) = 0.3 + 0.6 – (0.3)(0.6) = 0.72$. Note the $P(J \text{ AND } K) = P(J)P(K)$ because J and K are independent.

d. False – J and K are independent so $P(J) = P(J|K)$

46. a. $P(5)$

Chapter 7

47. a. $U(0, 4)$

48. b. 2 hour

49. a. $\frac{1}{4}$

50.

a. 0.7165

b. 4.16

c. 0

51. c. 5 years

52. c. exponential

53. 0.63

54. $B(14, 0.20)$

55. $B(14, 0.20)$

Chapter 8

56. c. the mean amount of weight lost by 15 people on the special weight loss diet.

57. 0.9951

58. 12.99

59. c. $\frac{1}{2}$

60. b. 0.60

61. c. $N(60, 5.477)$

62. 0.9990

63. a. eight days

64. c. 0.7500

65. a. 80%

66. b. 35%

67. b. no

68. b. quantitative continuous

69. c. 150

70. d. 0.06

71. c. 0.44

72. b. 0

Chapter 9

73. d. Matt is shorter than the average 14 year old boy.

74. Answers will vary.

75.

x	Relative Frequency	Cumulative Relative Frequency
1	0.3	0.3
2	0.2	0.2
4	0.4	0.4
5	0.1	0.1

Table A10

76.

a. 2.8

b. 1.48

c. 90%

77. $M = 3$; $Q_1 = 1$; $Q_3 = 4$

78. 1 and 4

79. d. $\frac{8}{70}$

80. c. $\frac{40}{70}$

81. a. $\frac{9}{19}$

82. b. false

83. b. false

84. b. false

85.

a. X = the number of pies Lee bakes every day.

b. $P(20)$

c. 0.1122

86. CI: (5.25, 8.48)

87.

a. uniform

b. exponential

c. normal

Chapter 10

88. $\frac{77}{100}$

89. $\frac{12}{42}$

90.

a. false
b. false
c. true
d. false

91. $N(180, 16.43)$

92. a. The distribution for $\bar{X}$ is still uniform with the same mean and standard deviation as the distribution for X.

93. c. The distribution for $\sum X$ is normal with a larger mean and a larger standard deviation than the distribution for X.

94. $N\left(2, \frac{0.25}{\sqrt{16}}\right)$

95. Answers will vary.

96. 0.5000

97. 7.6

98. 5

99. 0.9431

Chapter 11

100. 7.5

101. 0.0122

102. $N(7, 0.63)$

103. 0.9911

104. b. Exponential

105.

a. true
b. false
c. false

106. Answers will vary.

107. Student's t with $df = 15$

108. (560.07, 719.93)

109. quantitative continuous data

110. quantitative discrete data

111.

a. X = the number of patients with a shotgun wound the emergency room gets per 28 days
b. $P(4)$
c. 0.0183

112. greater than

113. No; $P(x = 8) = 0.0348$

114. You will lose $5.

115. Becca

116. 14

117. Sample mean = 3.2
Sample standard deviation = 1.85
Median = 3
$Q_1 = 2$
$Q_3 = 5$
$IQR = 3$

118. d. $z = –1.19$
e. 0.1171
f. Do not reject the null hypothesis.

119. We conclude that the patient does have the HIV virus when, in fact, the patient does not.

120. c. $z = 2.21$; $p = 0.0136$
d. Reject the null hypothesis.
e. We conclude that the proportion of Californian professionals that wear jeans to work is greater than the proportion of non-Californian professionals when, in fact, it is not greater.
f. We cannot conclude that the proportion of Californian professionals that wear jeans to work is greater than the proportion of non-Californian professionals when, in fact, it is greater.

121. c. dependent means

122. t_5

Chapter 12

123. (0.0424, 0.0770)

124. 2,401

125. Check student's solution.

126. 0.6321

127. $360

128. $N\left(72, \frac{72}{\sqrt{5}}\right)$

Chapter 13

129. 0.02

130. 0.40

131. $\frac{100}{140}$

132. $\frac{10}{60}$

133. p-value = 0; Reject the null hypothesis; conclude that they are dependent events

134. 8.4

135. $B(14, 0.60)$

136. d. Binomial

137. 0.3669

138. p-value = 0.0006; reject the null hypothesis; conclude that the averages are not equal

139. p-value = 0; reject the null hypothesis; conclude that the proportion of males is higher

140. Minimize α and β

141.

a. No

b. Yes, $P(M \text{ AND } 30+) = 0$

142. $\frac{12}{38}$

143. No; p-value = 0

144. a. uniform

References

Data from the *San Jose Mercury News*.

Baran, Daya. “20 Percent of Americans Have Never Used Email.” Webguild.org, 2010. Available online at: http://www.webguild.org/20080519/20-percent-of-americans-have-never-used-email (accessed October 17, 2013).

Data from *Parade Magazine*.

APPENDIX B: PRACTICE TESTS (1-4) AND FINAL EXAMS

Practice Test 1

1.1: Definitions of Statistics, Probability, and Key Terms

Use the following information to answer the next three exercises. A grocery store is interested in how much money, on average, their customers spend each visit in the produce department. Using their store records, they draw a sample of 1,000 visits and calculate each customer's average spending on produce.

1. Identify the population, sample, parameter, statistic, variable, and data for this example.

a. population
b. sample
c. parameter
d. statistic
e. variable
f. data

2. What kind of data is "amount of money spent on produce per visit"?

a. qualitative
b. quantitative-continuous
c. quantitative-discrete

3. The study finds that the mean amount spent on produce per visit by the customers in the sample is $12.84. This is an example of a:

a. population
b. sample
c. parameter
d. statistic
e. variable

1.2: Data, Sampling, and Variation in Data and Sampling

Use the following information to answer the next two exercises. A health club is interested in knowing how many times a typical member uses the club in a week. They decide to ask every tenth customer on a specified day to complete a short survey including information about how many times they have visited the club in the past week.

4. What kind of a sampling design is this?

a. cluster
b. stratified

c. simple random

d. systematic

5. "Number of visits per week" is what kind of data?

a. qualitative

b. quantitative-continuous

c. quantitative-discrete

6. Describe a situation in which you would calculate a parameter, rather than a statistic.

7. The U.S. federal government conducts a survey of high school seniors concerning their plans for future education and employment. One question asks whether they are planning to attend a four-year college or university in the following year. Fifty percent answer yes to this question; that fifty percent is a:

a. parameter

b. statistic

c. variable

d. data

8. Imagine that the U.S. federal government had the means to survey all high school seniors in the U.S. concerning their plans for future education and employment, and found that 50 percent were planning to attend a 4-year college or university in the following year. This 50 percent is an example of a:

a. parameter

b. statistic

c. variable

d. data

Use the following information to answer the next three exercises. A survey of a random sample of 100 nurses working at a large hospital asked how many years they had been working in the profession. Their answers are summarized in the following (incomplete) table.

9. Fill in the blanks in the table and round your answers to two decimal places for the Relative Frequency and Cumulative Relative Frequency cells.

# of years	Frequency	Relative Frequency	Cumulative Relative Frequency
< 5	25		
5–10	30		
> 10	empty		

Table B1

10. What proportion of nurses have five or more years of experience?

11. What proportion of nurses have ten or fewer years of experience?

12. Describe how you might draw a random sample of 30 students from a lecture class of 200 students.

13. Describe how you might draw a stratified sample of students from a college, where the strata are the students' class standing (freshman, sophomore, junior, or senior).

14. A manager wants to draw a sample, without replacement, of 30 employees from a workforce of 150. Describe how the chance of being selected will change over the course of drawing the sample.

15. The manager of a department store decides to measure employee satisfaction by selecting four departments at random, and conducting interviews with all the employees in those four departments. What type of survey design is this?

a. cluster

b. stratified
c. simple random
d. systematic

16. A popular American television sports program conducts a poll of viewers to see which team they believe will win the NFL (National Football League) championship this year. Viewers vote by calling a number displayed on the television screen and telling the operator which team they think will win. Do you think that those who participate in this poll are representative of all football fans in America?

17. Two researchers studying vaccination rates independently draw samples of 50 children, ages 3–18 months, from a large urban area, and determine if they are up to date on their vaccinations. One researcher finds that 84 percent of the children in her sample are up to date, and the other finds that 86 percent in his sample are up to date. Assuming both followed proper sampling procedures and did their calculations correctly, what is a likely explanation for this discrepancy?

18. A high school increased the length of the school day from 6.5 to 7.5 hours. Students who wished to attend this high school were required to sign contracts pledging to put forth their best effort on their school work and to obey the school rules; if they did not wish to do so, they could attend another high school in the district. At the end of one year, student performance on statewide tests had increased by ten percentage points over the previous year. Does this improvement prove that a longer school day improves student achievement?

19. You read a newspaper article reporting that eating almonds leads to increased life satisfaction. The study was conducted by the Almond Growers Association, and was based on a randomized survey asking people about their consumption of various foods, including almonds, and also about their satisfaction with different aspects of their life. Does anything about this poll lead you to question its conclusion?

20. Why is non-response a problem in surveys?

1.3: Frequency, Frequency Tables, and Levels of Measurement

21. Compute the mean of the following numbers, and report your answer using one more decimal place than is present in the original data:
14, 5, 18, 23, 6

1.4: Experimental Design and Ethics

22. A psychologist is interested in whether the size of tableware (bowls, plates, etc.) influences how much college students eat. He randomly assigns 100 college students to one of two groups: the first is served a meal using normal-sized tableware, while the second is served the same meal, but using tableware that it 20 percent smaller than normal. He records how much food is consumed by each group. Identify the following components of this study.

a. population
b. sample
c. experimental units
d. explanatory variable
e. treatment
f. response variable

23. A researcher analyzes the results of the SAT (Scholastic Aptitude Test) over a five-year period and finds that male students on average score higher on the math section, and female students on average score higher on the verbal section. She concludes that these observed differences in test performance are due to genetic factors. Explain how lurking variables could offer an alternative explanation for the observed differences in test scores.

24. Explain why it would not be possible to use random assignment to study the health effects of smoking.

25. A professor conducts a telephone survey of a city's population by drawing a sample of numbers from the phone book and having her student assistants call each of the selected numbers once to administer the survey. What are some sources of bias with this survey?

26. A professor offers extra credit to students who take part in her research studies. What is an ethical problem with this method of recruiting subjects?

2.1: Stem-and Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs

Use the following information to answer the next four exercises. The midterm grades on a chemistry exam, graded on a scale of 0 to 100, were:
62, 64, 65, 65, 68, 70, 72, 72, 74, 75, 75, 75, 76,78, 78, 81, 83, 83, 84, 85, 87, 88, 92, 95, 98, 98, 100, 100, 740

27. Do you see any outliers in this data? If so, how would you address the situation?

28. Construct a stem plot for this data, using only the values in the range 0–100.

29. Describe the distribution of exam scores.

2.2: Histograms, Frequency Polygons, and Time Series Graphs

30. In a class of 35 students, seven students received scores in the 70–79 range. What is the relative frequency of scores in this range?

Use the following information to answer the next three exercises. You conduct a poll of 30 students to see how many classes they are taking this term. Your results are:
1; 1; 1; 1
2; 2; 2; 2; 2
3; 3; 3; 3; 3; 3; 3; 3
4; 4; 4; 4; 4; 4; 4; 4; 4
5; 5; 5; 5

31. You decide to construct a histogram of this data. What will be the range of your first bar, and what will be the central point?

32. What will be the widths and central points of the other bars?

33. Which bar in this histogram will be the tallest, and what will be its height?

34. You get data from the U.S. Census Bureau on the median household income for your city, and decide to display it graphically. Which is the better choice for this data, a bar graph or a histogram?

35. You collect data on the color of cars driven by students in your statistics class, and want to display this information graphically. Which is the better choice for this data, a bar graph or a histogram?

2.3: Measures of the Location of the Data

36. Your daughter brings home test scores showing that she scored in the 80^{th} percentile in math and the 76^{th} percentile in reading for her grade. Interpret these scores.

37. You have to wait 90 minutes in the emergency room of a hospital before you can see a doctor. You learn that your wait time was in the 82^{nd} percentile of all wait times. Explain what this means, and whether you think it is good or bad.

2.4: Box Plots

Use the following information to answer the next three exercises. 1; 1; 2; 3; 4; 4; 5; 5; 6; 7; 7; 8; 9

38. What is the median for this data?

39. What is the first quartile for this data?

40. What is the third quartile for this data?

Use the following information to answer the next four exercises. This box plot represents scores on the final exam for a physics class.

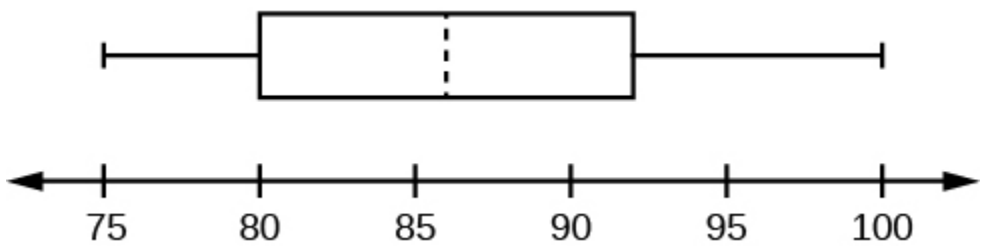

Figure B1

41. What is the median for this data, and how do you know?

42. What are the first and third quartiles for this data, and how do you know?

43. What is the interquartile range for this data?

44. What is the range for this data?

2.5: Measures of the Center of the Data

45. In a marathon, the median finishing time was 3:35:04 (three hours, 35 minutes, and four seconds). You finished in 3:34:10. Interpret the meaning of the median time, and discuss your time in relation to it.

Use the following information to answer the next three exercises. The value, in thousands of dollars, for houses on a block, are: 45; 47; 47.5; 51; 53.5; 125.

46. Calculate the mean for this data.

47. Calculate the median for this data.

48. Which do you think better reflects the average value of the homes on this block?

2.6: Skewness and the Mean, Median, and Mode

49. In a left-skewed distribution, which is greater?

a. the mean
b. the media
c. the mode

50. In a right-skewed distribution, which is greater?

a. the mean
b. the median
c. the mode

51. In a symmetrical distribution what will be the relationship among the mean, median, and mode?

2.7: Measures of the Spread of the Data

Use the following information to answer the next four exercises. 10; 11; 15; 15; 17; 22

52. Compute the mean and standard deviation for this data; use the sample formula for the standard deviation.

53. What number is two standard deviations above the mean of this data?

54. Express the number 13.7 in terms of the mean and standard deviation of this data.

55. In a biology class, the scores on the final exam were normally distributed, with a mean of 85, and a standard deviation of five. Susan got a final exam score of 95. Express her exam result as a *z*-score, and interpret its meaning.

3.1: Terminology

Use the following information to answer the next two exercises. You have a jar full of marbles: 50 are red, 25 are blue, and 15 are yellow. Assume you draw one marble at random for each trial, and replace it before the next trial.
Let $P(R)$ = the probability of drawing a red marble.
Let $P(B)$ = the probability of drawing a blue marble.
Let $P(Y)$ = the probability of drawing a yellow marble.

56. Find $P(B)$.

57. Which is more likely, drawing a red marble or a yellow marble? Justify your answer numerically.

Use the following information to answer the next two exercises. The following are probabilities describing a group of college students.
Let $P(M)$ = the probability that the student is male
Let $P(F)$ = the probability that the student is female
Let $P(E)$ = the probability the student is majoring in education
Let $P(S)$ = the probability the student is majoring in science

58. Write the symbols for the probability that a student, selected at random, is both female and a science major.

59. Write the symbols for the probability that the student is an education major, given that the student is male.

3.2: Independent and Mutually Exclusive Events

60. Events A and B are independent.
If $P(A) = 0.3$ and $P(B) = 0.5$, find $P(A \text{ AND } B)$.

61. C and D are mutually exclusive events.
If $P(C) = 0.18$ and $P(D) = 0.03$, find $P(C \text{ OR } D)$.

3.3: Two Basic Rules of Probability

62. In a high school graduating class of 300, 200 students are going to college, 40 are planning to work full-time, and 80 are taking a gap year. Are these events mutually exclusive?

Use the following information to answer the next two exercises. An archer hits the center of the target (the bullseye) 70 percent of the time. However, she is a streak shooter, and if she hits the center on one shot, her probability of hitting it on the shot immediately following is 0.85. Written in probability notation:
$P(A) = P(B) = P(\text{hitting the center on one shot}) = 0.70$
$P(B|A) = \text{P(hitting the center on a second shot, given that she hit it on the first)} = 0.85$

63. Calculate the probability that she will hit the center of the target on two consecutive shots.

64. Are $P(A)$ and $P(B)$ independent in this example?

3.4: Contingency Tables

Use the following information to answer the next three exercises. The following contingency table displays the number of students who report studying at least 15 hours per week, and how many made the honor roll in the past semester.

	Honor roll	No honor roll	Total
Study at least 15 hours/week		200	
Study less than 15 hours/week	125	193	
Total			1,000

Table B2

65. Complete the table.

66. Find $P(\text{honor roll}|\text{study at least 15 hours per week})$.

67. What is the probability a student studies less than 15 hours per week?

68. Are the events “study at least 15 hours per week” and “makes the honor roll” independent? Justify your answer numerically.

3.5: Tree and Venn Diagrams

69. At a high school, some students play on the tennis team, some play on the soccer team, but neither plays both tennis and soccer. Draw a Venn diagram illustrating this.

70. At a high school, some students play tennis, some play soccer, and some play both. Draw a Venn diagram illustrating this.

Practice Test 1 Solutions

1.1: Definitions of Statistics, Probability, and Key Terms

1.

a. population: all the shopping visits by all the store’s customers
b. sample: the 1,000 visits drawn for the study
c. parameter: the average expenditure on produce per visit by all the store’s customers
d. statistic: the average expenditure on produce per visit by the sample of 1,000
e. variable: the expenditure on produce for each visit
f. data: the dollar amounts spent on produce; for instance, $15.40, $11.53, etc

2. c

3. d

1.2: Data, Sampling, and Variation in Data and Sampling

4. d

5. c

6. Answers will vary.
Sample Answer: Any solution in which you use data from the entire population is acceptable. For instance, a professor might calculate the average exam score for her class: because the scores of all members of the class were used in the calculation, the average is a parameter.

7. b

8. a

9.

# of years	Frequency	Relative Frequency	Cumulative Relative Frequency
< 5	25	0.25	0.25
5–10	30	0.30	0.55
> 10	45	0.45	1.00

Table B3

10. 0.75

11. 0.55

12. Answers will vary.
Sample Answer: One possibility is to obtain the class roster and assign each student a number from 1 to 200. Then use a random number generator or table of random number to generate 30 numbers between 1 and 200, and select the students matching the random numbers. It would also be acceptable to write each student's name on a card, shuffle them in a box, and draw 30 names at random.

13. One possibility would be to obtain a roster of students enrolled in the college, including the class standing for each student. Then you would draw a proportionate random sample from within each class (for instance, if 30 percent of the students in the college are freshman, then 30 percent of your sample would be drawn from the freshman class).

14. For the first person picked, the chance of any individual being selected is one in 150. For the second person, it is one in 149, for the third it is one in 148, and so on. For the 30th person selected, the chance of selection is one in 121.

15. a

16. No. There are at least two chances for bias. First, the viewers of this particular program may not be representative of American football fans as a whole. Second, the sample will be self-selected, because people have to make a phone call in order to take part, and those people are probably not representative of the American football fan population as a whole.

17. These results (84 percent in one sample, 86 percent in the other) are probably due to sampling variability. Each researcher drew a different sample of children, and you would not expect them to get exactly the same result, although you would expect the results to be similar, as they are in this case.

18. No. The improvement could also be due to self-selection: only motivated students were willing to sign the contract, and they would have done well even in a school with 6.5 hour days. Because both changes were implemented at the same time, it is not possible to separate out their influence.

19. At least two aspects of this poll are troublesome. The first is that it was conducted by a group who would benefit by the result—almond sales are likely to increase if people believe that eating almonds will make them happier. The second is that this poll found that almond consumption and life satisfaction are correlated, but does not establish that eating almonds causes satisfaction. It is equally possible, for instance, that people with higher incomes are more likely to eat almonds, and are also more satisfied with their lives.

20. You want the sample of people who take part in a survey to be representative of the population from which they are drawn. People who refuse to take part in a survey often have different views than those who do participate, and so even a random sample may produce biased results if a large percentage of those selected refuse to participate in a survey.

1.3: Frequency, Frequency Tables, and Levels of Measurement

21. 13.2

1.4: Experimental Design and Ethics

22.

a. population: all college students
b. sample: the 100 college students in the study
c. experimental units: each individual college student who participated
d. explanatory variable: the size of the tableware
e. treatment: tableware that is 20 percent smaller than normal
f. response variable: the amount of food eaten

23. There are many lurking variables that could influence the observed differences in test scores. Perhaps the boys, on average, have taken more math courses than the girls, and the girls have taken more English classes than the boys. Perhaps the boys have been encouraged by their families and teachers to prepare for a career in math and science, and thus have put more effort into studying math, while the girls have been encouraged to prepare for fields like communication and psychology that are more focused on language use. A study design would have to control for these and other potential lurking variables (anything that could explain the observed difference in test scores, other than the genetic explanation) in order to draw a scientifically sound conclusion about genetic differences.

24. To use random assignment, you would have to be able to assign people to either smoke or not smoke. Because smoking has many harmful effects, this would not be an ethical experiment. Instead, we study people who have chosen to smoke, and compare them to others who have chosen not to smoke, and try to control for the other ways those two groups may differ (lurking variables).

25. Sources of bias include the fact that not everyone has a telephone, that cell phone numbers are often not listed in published directories, and that an individual might not be at home at the time of the phone call; all these factors make it likely that the respondents to the survey will not be representative of the population as a whole.

26. Research subjects should not be coerced into participation, and offering extra credit in exchange for participation could be construed as coercion. In addition, this method will result in a volunteer sample, which cannot be assumed to be representative of the population as a whole.

2.1: Stem-and Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs

27. The value 740 is an outlier, because the exams were graded on a scale of 0 to 100, and 740 is far outside that range. It may be a data entry error, with the actual score being 74, so the professor should check that exam again to see what the actual score was.

28.

Stem	Leaf
6	2 4 5 5 8
7	0 2 2 4 5 5 5 6 8 8
8	1 3 3 4 5 7 8
9	2 5 8 8
10	0 0

Table B4

29. Most scores on this exam were in the range of 70–89, with a few scoring in the 60–69 range, and a few in the 90–100 range.

2.2: Histograms, Frequency Polygons, and Time Series Graphs

30. $RF = \frac{7}{35} = 0.2$

31. The range will be 0.5–1.5, and the central point will be 1.

32. Range 1.5–2.5, central point 2; range 2.5–3.5, central point 3; range 3.5–4.5, central point 4; range 4.5–5.5., central point 5.

33. The bar from 3.5 to 4.5, with a central point of 4, will be tallest; its height will be nine, because there are nine students taking four courses.

34. The histogram is a better choice, because income is a continuous variable.

35. A bar graph is the better choice, because this data is categorical rather than continuous.

2.3: Measures of the Location of the Data

36. Your daughter scored better than 80 percent of the students in her grade on math and better than 76 percent of the students in reading. Both scores are very good, and place her in the upper quartile, but her math score is slightly better in relation to her peers than her reading score.

37. You had an unusually long wait time, which is bad: 82 percent of patients had a shorter wait time than you, and only 18 percent had a longer wait time.

2.4: Box Plots

38. 5

39. 3

40. 7

41. The median is 86, as represented by the vertical line in the box.

42. The first quartile is 80, and the third quartile is 92, as represented by the left and right boundaries of the box.

43. $IQR = 92 - 80 = 12$

44. Range = 100 – 75 = 25

2.5: Measures of the Center of the Data

45. Half the runners who finished the marathon ran a time faster than 3:35:04, and half ran a time slower than 3:35:04. Your time is faster than the median time, so you did better than more than half of the runners in this race.

46. 61.5, or $61,500

47. 49.25 or $49,250

48. The median, because the mean is distorted by the high value of one house.

2.6: Skewness and the Mean, Median, and Mode

49. c

50. a

51. They will all be fairly close to each other.

2.7: Measures of the Spread of the Data

52. Mean: 15
Standard deviation: 4.3

$$\mu = \frac{10 + 11 + 15 + 15 + 17 + 22}{6} = 15$$

$$s = \sqrt{\frac{\sum \left(x - \bar{x}\right)^2}{n - 1}} = \sqrt{\frac{94}{5}} = 4.3$$

53. 15 + (2)(4.3) = 23.6

54. 13.7 is one standard deviation below the mean of this data, because 15 – 4.3 = 10.7

55. $z = \frac{95 - 85}{5} = 2.0$

Susan's *z*-score was 2.0, meaning she scored two standard deviations above the class mean for the final exam.

3.1: Terminology

56. $P(B) = \frac{25}{90} = 0.28$

57. Drawing a red marble is more likely.

$P(R) = \frac{50}{80} = 0.62$

$P(Y) = \frac{15}{80} = 0.19$

58. $P(F \text{ AND } S)$

59. $P(E|M)$

3.2: Independent and Mutually Exclusive Events

60. $P(A \text{ AND } B) = (0.3)(0.5) = 0.15$

61. $P(C \text{ OR } D) = 0.18 + 0.03 = 0.21$

3.3: Two Basic Rules of Probability

62. No, they cannot be mutually exclusive, because they add up to more than 300. Therefore, some students must fit into two or more categories (e.g., both going to college and working full time).

63. $P(A \text{ and } B) = (P(B|A))(P(A)) = (0.85)(0.70) = 0.595$

64. No. If they were independent, $P(B)$ would be the same as $P(B|A)$. We know this is not the case, because $P(B) = 0.70$ and $P(B|A) = 0.85$.

3.4: Contingency Tables

65.

	Honor roll	No honor roll	Total
Study at least 15 hours/week	482	200	682
Study less than 15 hours/week	125	193	318
Total	607	393	1,000

Table B5

66. $P(\text{honor roll|study at least 15 hours word per week}) = \frac{482}{1000} = 0.482$

67. $P(\text{studies less than 15 hours word per week}) = \frac{125 + 193}{1000} = 0.318$

68. Let $P(S)$ = study at least 15 hours per week
Let $P(H)$ = makes the honor roll
From the table, $P(S) = 0.682$, $P(H) = 0.607$, and $P(S \text{ AND } H) = 0.482$.
If $P(S)$ and $P(H)$ were independent, then $P(S \text{ AND } H)$ would equal $(P(S))(P(H))$.
However, $(P(S))(P(H)) = (0.682)(0.607) = 0.414$, while $P(S \text{ AND } H) = 0.482$.
Therefore, $P(S)$ and $P(H)$ are not independent.

3.5: Tree and Venn Diagrams

69.

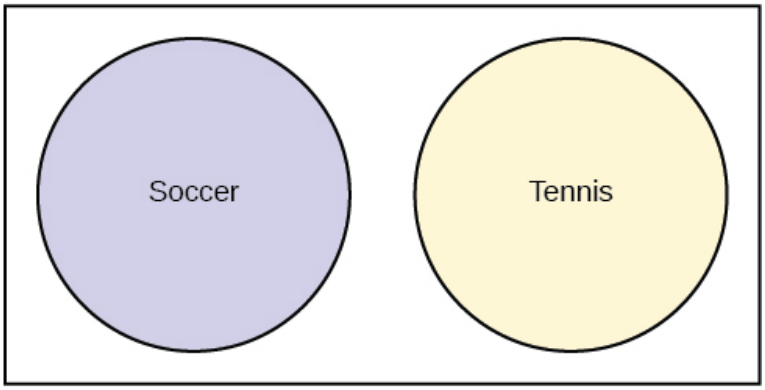

Figure B2

70.

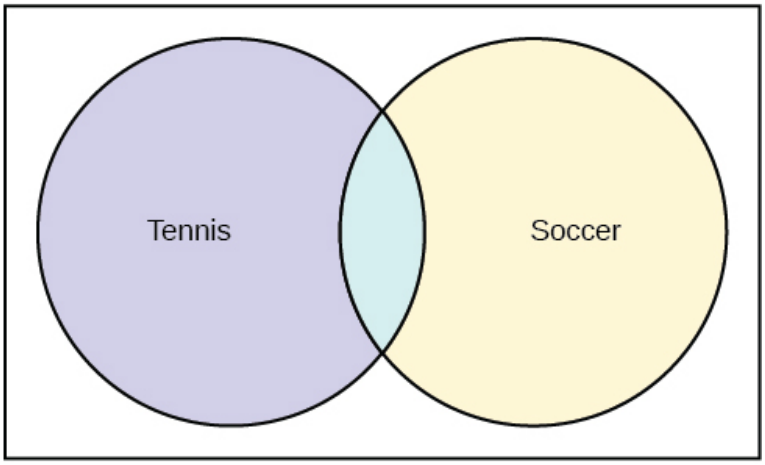

Figure B3

Practice Test 2

4.1: Probability Distribution Function (PDF) for a Discrete Random Variable

Use the following information to answer the next five exercises. You conduct a survey among a random sample of students at a particular university. The data collected includes their major, the number of classes they took the previous semester, and amount of money they spent on books purchased for classes in the previous semester.

1. If X = student's major, then what is the domain of X?

2. If Y = the number of classes taken in the previous semester, what is the domain of Y?

3. If Z = the amount of money spent on books in the previous semester, what is the domain of Z?

4. Why are X, Y, and Z in the previous example random variables?

5. After collecting data, you find that for one case, $z = –7$. Is this a possible value for Z?

6. What are the two essential characteristics of a discrete probability distribution?

Use this discrete probability distribution represented in this table to answer the following six questions. The university library records the number of books checked out by each patron over the course of one day, with the following result:

x	$P(x)$
0	0.20

Table B6

x	*P*(*x*)
1	0.45
2	0.20
3	0.10
4	0.05

Table B6

7. Define the random variable X for this example.

8. What is $P(x > 2)$?

9. What is the probability that a patron will check out at least one book?

10. What is the probability a patron will take out no more than three books?

11. If the table listed $P(x)$ as 0.15, how would you know that there was a mistake?

12. What is the average number of books taken out by a patron?

4.2: Mean or Expected Value and Standard Deviation

Use the following information to answer the next four exercises. Three jobs are open in a company: one in the accounting department, one in the human resources department, and one in the sales department. The accounting job receives 30 applicants, and the human resources and sales department 60 applicants.

13. If X = the number of applications for a job, use this information to fill in **Table B7**.

x	*P*(*x*)	*xP*(*x*)

Table B7

14. What is the mean number of applicants?

15. What is the PDF for X?

16. Add a fourth column to the table, for $(x - \mu)^2 P(x)$.

17. What is the standard deviation of X?

4.3: Binomial Distribution

18. In a binomial experiment, if $p = 0.65$, what does q equal?

19. What are the required characteristics of a binomial experiment?

20. Joe conducts an experiment to see how many times he has to flip a coin before he gets four heads in a row. Does this qualify as a binomial experiment?

Use the following information to answer the next three exercises. In a particularly community, 65 percent of households include at least one person who has graduated from college. You randomly sample 100 households in this community. Let X = the number of households including at least one college graduate.

21. Describe the probability distribution of X.

22. What is the mean of X?

23. What is the standard deviation of X?

Use the following information to answer the next four exercises. Joe is the star of his school's baseball team. His batting average is 0.400, meaning that for every ten times he comes to bat (an at-bat), four of those times he gets a hit. You decide to track his batting performance his next 20 at-bats.

24. Define the random variable X in this experiment.

25. Assuming Joe's probability of getting a hit is independent and identical across all 20 at-bats, describe the distribution of X.

26. Given this information, what number of hits do you predict Joe will get?

27. What is the standard deviation of X?

4.4: Geometric Distribution

28. What are the three major characteristics of a geometric experiment?

29. You decide to conduct a geometric experiment by flipping a coin until it comes up heads. This takes five trials. Represent the outcomes of this trial, using H for heads and T for tails.

30. You are conducting a geometric experiment by drawing cards from a normal 52-card pack, with replacement, until you draw the Queen of Hearts. What is the domain of X for this experiment?

31. You are conducting a geometric experiment by drawing cards from a normal 52-card deck, without replacement, until you draw a red card. What is the domain of X for this experiment?

Use the following information to answer the next three exercises. In a particular university, 27 percent of students are engineering majors. You decide to select students at random until you choose one that is an engineering major. Let X = the number of students you select until you find one that is an engineering major.

32. What is the probability distribution of X?

33. What is the mean of X?

34. What is the standard deviation of X?

4.5: Hypergeometric Distribution

35. You draw a random sample of ten students to participate in a survey, from a group of 30, consisting of 16 boys and 14 girls. You are interested in the probability that seven of the students chosen will be boys. Does this qualify as a hypergeometric experiment? List the conditions and whether or not they are met.

36. You draw five cards, without replacement, from a normal 52-card deck of playing cards, and are interested in the probability that two of the cards are spades. What are the group of interest, size of the group of interest, and sample size for this example?

4.6: Poisson Distribution

37. What are the key characteristics of the Poisson distribution?

Use the following information to answer the next three exercises. The number of drivers to arrive at a toll booth in an hour can be modeled by the Poisson distribution.

38. If X = the number of drivers, and the average numbers of drivers per hour is four, how would you express this distribution?

39. What is the domain of X?

40. What are the mean and standard deviation of X?

5.1: Continuous Probability Functions

41. You conduct a survey of students to see how many books they purchased the previous semester, the total amount they paid for those books, the number they sold after the semester was over, and the amount of money they received for the books they sold. Which variables in this survey are discrete, and which are continuous?

42. With continuous random variables, we never calculate the probability that X has a particular value, but always speak in terms of the probability that X has a value within a particular range. Why is this?

43. For a continuous random variable, why are $P(x < c)$ and $P(x \leq c)$ equivalent statements?

44. For a continuous probability function, $P(x < 5) = 0.35$. What is $P(x > 5)$, and how do you know?

45. Describe how you would draw the continuous probability distribution described by the function $f(x) = \frac{1}{10}$ for $0 \leq x \leq 10$. What type of a distribution is this?

46. For the continuous probability distribution described by the function $f(x) = \frac{1}{10}$ for $0 \leq x \leq 10$, what is the $P(0 < x < 4)$?

5.2: The Uniform Distribution

47. For the continuous probability distribution described by the function $f(x) = \frac{1}{10}$ for $0 \leq x \leq 10$, what is the $P(2 < x < 5)$?

Use the following information to answer the next four exercises. The number of minutes that a patient waits at a medical clinic to see a doctor is represented by a uniform distribution between zero and 30 minutes, inclusive.

48. If X equals the number of minutes a person waits, what is the distribution of X?

49. Write the probability density function for this distribution.

50. What is the mean and standard deviation for waiting time?

51. What is the probability that a patient waits less than ten minutes?

5.3: The Exponential Distribution

52. The distribution of the variable X, representing the average time to failure for an automobile battery, can be written as: $X \sim Exp(m)$. Describe this distribution in words.

53. If the value of m for an exponential distribution is ten, what are the mean and standard deviation for the distribution?

54. Write the probability density function for a variable distributed as: $X \sim Exp(0.2)$.

6.1: The Standard Normal Distribution

55. Translate this statement about the distribution of a random variable X into words: $X \sim (100, 15)$.

56. If the variable X has the standard normal distribution, express this symbolically.

Use the following information for the next six exercises. According to the World Health Organization, distribution of height in centimeters for girls aged five years and no months has the distribution: $X \sim N(109, 4.5)$.

57. What is the z-score for a height of 112 inches?

58. What is the z-score for a height of 100 centimeters?

59. Find the z-score for a height of 105 centimeters and explain what that means In the context of the population.

60. What height corresponds to a z-score of 1.5 in this population?

61. Using the empirical rule, we expect about 68 percent of the values in a normal distribution to lie within one standard deviation above or below the mean. What does this mean, in terms of a specific range of values, for this distribution?

62. Using the empirical rule, about what percent of heights in this distribution do you expect to be between 95.5 cm and 122.5 cm?

6.2: Using the Normal Distribution

Use the following information to answer the next four exercises. The distributor of lotto tickets claims that 20 percent of the tickets are winners. You draw a sample of 500 tickets to test this proposition.

63. Can you use the normal approximation to the binomial for your calculations? Why or why not.

64. What are the expected mean and standard deviation for your sample, assuming the distributor's claim is true?

65. What is the probability that your sample will have a mean greater than 100?

66. If the z-score for your sample result is –2.00, explain what this means, using the empirical rule.

7.1: The Central Limit Theorem for Sample Means (Averages)

67. What does the central limit theorem state with regard to the distribution of sample means?

68. The distribution of results from flipping a fair coin is uniform: heads and tails are equally likely on any flip, and over a large number of trials, you expect about the same number of heads and tails. Yet if you conduct a study by flipping 30 coins and recording the number of heads, and repeat this 100 times, the distribution of the mean number of heads will be approximately normal. How is this possible?

69. The mean of a normally-distributed population is 50, and the standard deviation is four. If you draw 100 samples of size 40 from this population, describe what you would expect to see in terms of the sampling distribution of the sample mean.

70. X is a random variable with a mean of 25 and a standard deviation of two. Write the distribution for the sample mean of samples of size 100 drawn from this population.

71. Your friend is doing an experiment drawing samples of size 50 from a population with a mean of 117 and a standard deviation of 16. This sample size is large enough to allow use of the central limit theorem, so he says the standard deviation of the sampling distribution of sample means will also be 16. Explain why this is wrong, and calculate the correct value.

72. You are reading a research article that refers to "the standard error of the mean." What does this mean, and how is it calculated?

Use the following information to answer the next six exercises. You repeatedly draw samples of $n = 100$ from a population with a mean of 75 and a standard deviation of 4.5.

73. What is the expected distribution of the sample means?

74. One of your friends tries to convince you that the standard error of the mean should be 4.5. Explain what error your friend made.

75. What is the *z*-score for a sample mean of 76?

76. What is the *z*-score for a sample mean of 74.7?

77. What sample mean corresponds to a *z*-score of 1.5?

78. If you decrease the sample size to 50, will the standard error of the mean be smaller or larger? What would be its value?

Use the following information to answer the next two questions. We use the empirical rule to analyze data for samples of size 60 drawn from a population with a mean of 70 and a standard deviation of 9.

79. What range of values would you expect to include 68 percent of the sample means?

80. If you increased the sample size to 100, what range would you expect to contain 68 percent of the sample means, applying the empirical rule?

7.2: The Central Limit Theorem for Sums

81. How does the central limit theorem apply to sums of random variables?

82. Explain how the rules applying the central limit theorem to sample means, and to sums of a random variable, are similar.

83. If you repeatedly draw samples of size 50 from a population with a mean of 80 and a standard deviation of four, and calculate the sum of each sample, what is the expected distribution of these sums?

Use the following information to answer the next four exercises. You draw one sample of size 40 from a population with a mean of 125 and a standard deviation of seven.

84. Compute the sum. What is the probability that the sum for your sample will be less than 5,000?

85. If you drew samples of this size repeatedly, computing the sum each time, what range of values would you expect to contain 95 percent of the sample sums?

86. What value is one standard deviation below the mean?

87. What value corresponds to a *z*-score of 2.2?

7.3: Using the Central Limit Theorem

88. What does the law of large numbers say about the relationship between the sample mean and the population mean?

89. Applying the law of large numbers, which sample mean would expect to be closer to the population mean, a sample of size ten or a sample of size 100?

Use this information for the next three questions. A manufacturer makes screws with a mean diameter of 0.15 cm (centimeters) and a range of 0.10 cm to 0.20 cm; within that range, the distribution is uniform.

90. If X = the diameter of one screw, what is the distribution of X?

91. Suppose you repeatedly draw samples of size 100 and calculate their mean. Applying the central limit theorem, what is the distribution of these sample means?

92. Suppose you repeatedly draw samples of 60 and calculate their sum. Applying the central limit theorem, what is the distribution of these sample sums?

Practice Test 2 Solutions

Probability Distribution Function (PDF) for a Discrete Random Variable

1. The domain of X = {English, Mathematics,….], i.e., a list of all the majors offered at the university, plus "undeclared."

2. The domain of Y = {0, 1, 2, …}, i.e., the integers from 0 to the upper limit of classes allowed by the university.

3. The domain of Z = any amount of money from 0 upwards.

4. Because they can take any value within their domain, and their value for any particular case is not known until the survey is completed.

5. No, because the domain of Z includes only positive numbers (you can't spend a negative amount of money). Possibly the value –7 is a data entry error, or a special code to indicated that the student did not answer the question.

6. The probabilities must sum to 1.0, and the probabilities of each event must be between 0 and 1, inclusive.

7. Let X = the number of books checked out by a patron.

8. $P(x > 2) = 0.10 + 0.05 = 0.15$

9. $P(x \geq 0) = 1 - 0.20 = 0.80$

10. $P(x \leq 3) = 1 - 0.05 = 0.95$

11. The probabilities would sum to 1.10, and the total probability in a distribution must always equal 1.0.

12. $\bar{x} = 0(0.20) + 1(0.45) + 2(0.20) + 3(0.10) + 4(0.05) = 1.35$

Mean or Expected Value and Standard Deviation

13.

x	$P(x)$	$xP(x)$
30	0.33	9.90
40	0.33	13.20
60	0.33	19.80

Table B8

14. $\bar{x} = 9.90 + 13.20 + 19.80 = 42.90$

15. $P(x = 30) = 0.33$
$P(x = 40) = 0.33$
$P(x = 60) = 0.33$

16.

x	$P(x)$	$xP(x)$	$(x - \mu)^2P(x)$
30	0.33	9.90	$(30 - 42.90)^2(0.33) = 54.91$
40	0.33	13.20	$(40 - 42.90)^2(0.33) = 2.78$
60	0.33	19.90	$(60 - 42.90)^2(0.33) = 96.49$

Table B9

17. $\sigma_x = \sqrt{54.91 + 2.78 + 96.49} = 12.42$

Binomial Distribution

18. $q = 1 - 0.65 = 0.35$

19.

1. There are a fixed number of trials.
2. There are only two possible outcomes, and they add up to 1.
3. The trials are independent and conducted under identical conditions.

20. No, because there are not a fixed number of trials

21. $X \sim B(100, 0.65)$

22. $\mu = np = 100(0.65) = 65$

23. $\sigma_x = \sqrt{npq} = \sqrt{100(0.65)(0.35)} = 4.77$

24. X = Joe gets a hit in one at-bat (in one occasion of his coming to bat)

25. $X \sim B(20, 0.4)$

26. $\mu = np = 20(0.4) = 8$

27. $\sigma_x = \sqrt{npq} = \sqrt{20(0.40)(0.60)} = 2.19$

4.4: Geometric Distribution

28.

1. A series of Bernoulli trials are conducted until one is a success, and then the experiment stops.
2. At least one trial is conducted, but there is no upper limit to the number of trials.
3. The probability of success or failure is the same for each trial.

29. *T T T T H*

30. The domain of $X = \{1, 2, 3, 4, 5, ….n\}$. Because you are drawing with replacement, there is no upper bound to the number of draws that may be necessary.

31. The domain of $X = \{1, 2, 3, 4, 5, 6, 7, 8., 9, 10, 11, 12…27\}$. Because you are drawing without replacement, and 26 of the 52 cards are red, you have to draw a red card within the first 17 draws.

32. $X \sim G(0.24)$

33. $\mu = \frac{1}{p} = \frac{1}{0.27} = 3.70$

34. $\sigma = \sqrt{\frac{1-p}{p^2}} = \sqrt{\frac{1-0.27}{0.27^2}} = 3.16$

4.5: Hypergeometric Distribution

35. Yes, because you are sampling from a population composed of two groups (boys and girls), have a group of interest (boys), and are sampling without replacement (hence, the probabilities change with each pick, and you are not performing Bernoulli trials).

36. The group of interest is the cards that are spades, the size of the group of interest is 13, and the sample size is five.

4.6: Poisson Distribution

37. A Poisson distribution models the number of events occurring in a fixed interval of time or space, when the events are independent and the average rate of the events is known.

38. $X \sim P(4)$

39. The domain of $X = \{0, 1, 2, 3, …..\}$ i.e., any integer from 0 upwards.

40. $\mu = 4$

$\sigma = \sqrt{4} = 2$

5.1: Continuous Probability Functions

41. The discrete variables are the number of books purchased, and the number of books sold after the end of the semester. The continuous variables are the amount of money spent for the books, and the amount of money received when they were sold.

42. Because for a continuous random variable, $P(x = c) = 0$, where c is any single value. Instead, we calculate $P(c < x < d)$, i.e., the probability that the value of x is between the values c and d.

43. Because $P(x = c) = 0$ for any continuous random variable.

44. $P(x > 5) = 1 – 0.35 = 0.65$, because the total probability of a continuous probability function is always 1.

45. This is a uniform probability distribution. You would draw it as a rectangle with the vertical sides at 0 and 20, and the horizontal sides at $\frac{1}{10}$ and 0.

46. $P(0 < x < 4) = (4 - 0)\left(\frac{1}{10}\right) = 0.4$

5.2: The Uniform Distribution

47. $P(2 < x < 5) = (5 - 2)\left(\frac{1}{10}\right) = 0.3$

48. $X \sim U(0, 15)$

49. $f(x) = \frac{1}{b-a}$ for $(a \leq x \leq b)$ so $f(x) = \frac{1}{30}$ for $(0 \leq x \leq 30)$

50. $\mu = \frac{a+b}{2} = \frac{0+30}{5} = 15.0$

$$\sigma = \sqrt{\frac{(b-a)^2}{12}} = \sqrt{\frac{(30-0)^2}{12}} = 8.66$$

51. $P(x < 10) = (10)\left(\frac{1}{30}\right) = 0.33$

5.3: The Exponential Distribution

52. X has an exponential distribution with decay parameter m and mean and standard deviation $\frac{1}{m}$. In this distribution, there will be a relatively large numbers of small values, with values becoming less common as they become larger.

53. $\mu = \sigma = \frac{1}{m} = \frac{1}{10} = 0.1$

54. $f(x) = 0.2e^{-0.2x}$ where $x \geq 0$.

6.1: The Standard Normal Distribution

55. The random variable X has a normal distribution with a mean of 100 and a standard deviation of 15.

56. $X \sim N(0,1)$

57. $z = \frac{x-\mu}{\sigma}$ so $z = \frac{112-109}{4.5} = 0.67$

58. $z = \frac{x-\mu}{\sigma}$ so $z = \frac{100-109}{4.5} = -2.00$

59. $z = \frac{105-109}{4.5} = -0.89$

This girl is shorter than average for her age, by 0.89 standard deviations.

60. 109 + (1.5)(4.5) = 115.75 cm

61. We expect about 68 percent of the heights of girls of age five years and zero months to be between 104.5 cm and 113.5 cm.

62. We expect 99.7 percent of the heights in this distribution to be between 95.5 cm and 122.5 cm, because that range represents the values three standard deviations above and below the mean.

6.2: Using the Normal Distribution

63. Yes, because both np and nq are greater than five.
$np = (500)(0.20) = 100$ and $nq = 500(0.80) = 400$

64. $\mu = np = (500)(0.20) = 100$

$$\sigma = \sqrt{npq} = \sqrt{500(0.20)(0.80)} = 8.94$$

65. Fifty percent, because in a normal distribution, half the values lie above the mean.

66. The results of our sample were two standard deviations below the mean, suggesting it is unlikely that 20 percent of the lotto tickets are winners, as claimed by the distributor, and that the true percent of winners is lower. Applying the Empirical Rule, If that claim were true, we would expect to see a result this far below the mean only about 2.5 percent of the time.

7.1: The Central Limit Theorem for Sample Means (Averages)

67. The central limit theorem states that if samples of sufficient size drawn from a population, the distribution of sample means will be normal, even if the distribution of the population is not normal.

68. The sample size of 30 is sufficiently large in this example to apply the central limit theorem. This theorem] states that for samples of sufficient size drawn from a population, the sampling distribution of the sample mean will approach normality, regardless of the distribution of the population from which the samples were drawn.

69. You would not expect each sample to have a mean of 50, because of sampling variability. However, you would expect the sampling distribution of the sample means to cluster around 50, with an approximately normal distribution, so that values close to 50 are more common than values further removed from 50.

70. $\bar{X} \sim N(25,\ 0.2)$ because $\bar{X} \sim N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right)$

71. The standard deviation of the sampling distribution of the sample means can be calculated using the formula $\left(\frac{\sigma_x}{\sqrt{n}}\right)$, which in this case is $\left(\frac{16}{\sqrt{50}}\right)$. The correct value for the standard deviation of the sampling distribution of the sample means is therefore 2.26.

72. The standard error of the mean is another name for the standard deviation of the sampling distribution of the sample mean. Given samples of size n drawn from a population with standard deviation σ_x, the standard error of the mean is $\left(\frac{\sigma_x}{\sqrt{n}}\right)$.

73. $X \sim N(75, 0.45)$

74. Your friend forgot to divide the standard deviation by the square root of *n*.

75. $z = \frac{\bar{x} - \mu_x}{\sigma_x} = \frac{76 - 75}{4.5} = 2.2$

76. $z = \frac{\bar{x} - \mu_x}{\sigma_x} = \frac{74.7 - 75}{4.5} = -0.67$

77. 75 + (1.5)(0.45) = 75.675

78. The standard error of the mean will be larger, because you will be dividing by a smaller number. The standard error of the mean for samples of size $n = 50$ is:
$\left(\frac{\sigma_x}{\sqrt{n}}\right) = \frac{4.5}{\sqrt{50}} = 0.64$

79. You would expect this range to include values up to one standard deviation above or below the mean of the sample means. In this case:
$70 + \frac{9}{\sqrt{60}} = 71.16$ and $70 - \frac{9}{\sqrt{60}} = 68.84$ so you would expect 68 percent of the sample means to be between 68.84 and 71.16.

80. $70 + \frac{9}{\sqrt{100}} = 70.9$ and $70 - \frac{9}{\sqrt{100}} = 69.1$ so you would expect 68 percent of the sample means to be between 69.1 and 70.9. Note that this is a narrower interval due to the increased sample size.

7.2: The Central Limit Theorem for Sums

81. For a random variable *X*, the random variable ΣX will tend to become normally distributed as the size n of the samples used to compute the sum increases.

82. Both rules state that the distribution of a quantity (the mean or the sum) calculated on samples drawn from a population will tend to have a normal distribution, as the sample size increases, regardless of the distribution of population from which the samples are drawn.

83. $\Sigma X \sim N(n\mu_x,\ (\sqrt{n})(\sigma_x))$ so $\Sigma X \sim N(4000,\ 28.3)$

84.The probability is 0.50, because 5,000 is the mean of the sampling distribution of sums of size 40 from this population. Sums of random variables computed from a sample of sufficient size are normally distributed, and in a normal distribution, half the values lie below the mean.

85. Using the empirical rule, you would expect 95 percent of the values to be within two standard deviations of the mean. Using the formula for the standard deviation is for a sample sum: $(\sqrt{n})(\sigma_x) = (\sqrt{40})(7) = 44.3$ so you would expect 95 percent of the values to be between 5,000 + (2)(44.3) and 5,000 – (2)(44.3), or between 4,911.4 and 588.6.

86. $\mu - (\sqrt{n})(\sigma_x) = 5000 - (\sqrt{40})(7) = 4955.7$

87. $5000 + (2.2)(\sqrt{40})(7) = 5097.4$

7.3: Using the Central Limit Theorem

88. The law of large numbers says that as sample size increases, the sample mean tends to get nearer and nearer to the population mean.

89. You would expect the mean from a sample of size 100 to be nearer to the population mean, because the law of large numbers says that as sample size increases, the sample mean tends to approach the population mea.

90. $X \sim N(0.10, 0.20)$

91. $\bar{X} \sim N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right)$ and the standard deviation of a uniform distribution is $\frac{b-a}{\sqrt{12}}$. In this example, the standard deviation of the distribution is $\frac{b-a}{\sqrt{12}} = \frac{0.10}{\sqrt{12}} = 0.03$

so $\bar{X} \sim N(0.15, 0.003)$

92. $\Sigma X \sim N((n)(\mu_x), (\sqrt{n})(\sigma_x))$ so $\Sigma X \sim N(9.0, 0.23)$

Practice Test 3

8.1: Confidence Interval, Single Population Mean, Population Standard Deviation Known, Normal

Use the following information to answer the next seven exercises. You draw a sample of size 30 from a normally distributed population with a standard deviation of four.

1. What is the standard error of the sample mean in this scenario, rounded to two decimal places?

2. What is the distribution of the sample mean?

3. If you want to construct a two-sided 95% confidence interval, how much probability will be in each tail of the distribution?

4. What is the appropriate *z*-score and error bound or margin of error (*EBM*) for a 95% confidence interval for this data?

5. Rounding to two decimal places, what is the 95% confidence interval if the sample mean is 41?

6. What is the 90% confidence interval if the sample mean is 41? Round to two decimal places

7. Suppose the sample size in this study had been 50, rather than 30. What would the 95% confidence interval be if the sample mean is 41? Round your answer to two decimal places.

8. For any given data set and sampling situation, which would you expect to be wider: a 95% confidence interval or a 99% confidence interval?

8.2: Confidence Interval, Single Population Mean, Standard Deviation Unknown, Student's *t*

9. Comparing graphs of the standard normal distribution (*z*-distribution) and a *t*-distribution with 15 degrees of freedom (*df*), how do they differ?

10. Comparing graphs of the standard normal distribution (*z*-distribution) and a *t*-distribution with 15 degrees of freedom (*df*), how are they similar?

Use the following information to answer the next five exercises. Body temperature is known to be distributed normally among healthy adults. Because you do not know the population standard deviation, you use the t-distribution to study body temperature. You collect data from a random sample of 20 healthy adults and find that your sample temperatures have a mean of 98.4 and a sample standard deviation of 0.3 (both in degrees Fahrenheit).

11. What is the degrees of freedom (*df*) for this study?

12. For a two-tailed 95% confidence interval, what is the appropriate *t*-value to use in the formula?

13. What is the 95% confidence interval?

14. What is the 99% confidence interval? Round to two decimal places.

15. Suppose your sample size had been 30 rather than 20. What would the 95% confidence interval be then? Round to two decimal places

8.3: Confidence Interval for a Population Proportion

Use this information to answer the next four exercises. You conduct a poll of 500 randomly selected city residents, asking them if they own an automobile. 280 say they do own an automobile, and 220 say they do not.

16. Find the sample proportion and sample standard deviation for this data.

17. What is the 95% two-sided confidence interval? Round to four decimal places.

18. Calculate the 90% confidence interval. Round to four decimal places.

19. Calculate the 99% confidence interval. Round to four decimal places.

Use the following information to answer the next three exercises. You are planning to conduct a poll of community members age 65 and older, to determine how many own mobile phones. You want to produce an estimate whose 95% confidence interval will be within four percentage points (plus or minus) the true population proportion. Use an estimated population proportion of 0.5.

20. What sample size do you need?

21. Suppose you knew from prior research that the population proportion was 0.6. What sample size would you need?

22. Suppose you wanted a 95% confidence interval within three percentage points of the population. Assume the population proportion is 0.5. What sample size do you need?

9.1: Null and Alternate Hypotheses

23. In your state, 58 percent of registered voters in a community are registered as Republicans. You want to conduct a study to see if this also holds up in your community. State the null and alternative hypotheses to test this.

24. You believe that at least 58 percent of registered voters in a community are registered as Republicans. State the null and alternative hypotheses to test this.

25. The mean household value in a city is $268,000. You believe that the mean household value in a particular neighborhood is lower than the city average. Write the null and alternative hypotheses to test this.

26. State the appropriate alternative hypothesis to this null hypothesis: H_0: $\mu = 107$

27. State the appropriate alternative hypothesis to this null hypothesis: H_0: $p < 0.25$

9.2: Outcomes and the Type I and Type II Errors

28. If you reject H_0 when H_0 is correct, what type of error is this?

29. If you fail to reject H_0 when H_0 is false, what type of error is this?

30. What is the relationship between the Type II error and the power of a test?

31. A new blood test is being developed to screen patients for cancer. Positive results are followed up by a more accurate (and expensive) test. It is assumed that the patient does not have cancer. Describe the null hypothesis, the Type I and Type II errors for this situation, and explain which type of error is more serious.

32. Explain in words what it means that a screening test for TB has an α level of 0.10. The null hypothesis is that the patient does not have TB.

33. Explain in words what it means that a screening test for TB has a β level of 0.20. The null hypothesis is that the patient does not have TB.

34. Explain in words what it means that a screening test for TB has a power of 0.80.

9.3: Distribution Needed for Hypothesis Testing

35. If you are conducting a hypothesis test of a single population mean, and you do not know the population variance, what test will you use if the sample size is 10 and the population is normal?

36. If you are conducting a hypothesis test of a single population mean, and you know the population variance, what test will you use?

37. If you are conducting a hypothesis test of a single population proportion, with np and nq greater than or equal to five, what test will you use, and with what parameters?

38. Published information indicates that, on average, college students spend less than 20 hours studying per week. You draw a sample of 25 students from your college, and find the sample mean to be 18.5 hours, with a standard deviation of 1.5 hours. What distribution will you use to test whether study habits at your college are the same as the national average, and why?

39. A published study says that 95 percent of American children are vaccinated against measles, with a standard deviation of 1.5 percent. You draw a sample of 100 children from your community and check their vaccination records, to see if the vaccination rate in your community is the same as the national average. What distribution will you use for this test, and why?

9.4: Rare Events, the Sample, Decision, and Conclusion

40. You are conducting a study with an α level of 0.05. If you get a result with a p-value of 0.07, what will be your decision?

41. You are conducting a study with $\alpha = 0.01$. If you get a result with a p-value of 0.006, what will be your decision?

Use the following information to answer the next five exercises. According to the World Health Organization, the average height of a one-year-old child is 29". You believe children with a particular disease are smaller than average, so you draw a sample of 20 children with this disease and find a mean height of 27.5" and a sample standard deviation of 1.5".

42. What are the null and alternative hypotheses for this study?

43. What distribution will you use to test your hypothesis, and why?

44. What is the test statistic and the p-value?

45. Based on your sample results, what is your decision?

46. Suppose the mean for your sample was 25.0. Redo the calculations and describe what your decision would be.

9.5: Additional Information and Full Hypothesis Test Examples

47. You conduct a study using $\alpha = 0.05$. What is the level of significance for this study?

48. You conduct a study, based on a sample drawn from a normally distributed population with a known variance, with the following hypotheses:
H_0: $\mu = 35.5$
H_a: $\mu \neq 35.5$
Will you conduct a one-tailed or two-tailed test?

49. You conduct a study, based on a sample drawn from a normally distributed population with a known variance, with the following hypotheses:
H_0: $\mu \geq 35.5$
H_a: $\mu < 35.5$
Will you conduct a one-tailed or two-tailed test?

Use the following information to answer the next three exercises. Nationally, 80 percent of adults own an automobile. You are interested in whether the same proportion in your community own cars. You draw a sample of 100 and find that 75 percent own cars.

50. What are the null and alternative hypotheses for this study?

51. What test will you use, and why?

10.1: Comparing Two Independent Population Means with Unknown Population Standard Deviations

52. You conduct a poll of political opinions, interviewing both members of 50 married couples. Are the groups in this study independent or matched?

53. You are testing a new drug to treat insomnia. You randomly assign 80 volunteer subjects to either the experimental (new drug) or control (standard treatment) conditions. Are the groups in this study independent or matched?

54. You are investigating the effectiveness of a new math textbook for high school students. You administer a pretest to a group of students at the beginning of the semester, and a posttest at the end of a year's instruction using this textbook, and compare the results. Are the groups in this study independent or matched?

Use the following information to answer the next two exercises. You are conducting a study of the difference in time at two colleges for undergraduate degree completion. At College A, students take an average of 4.8 years to complete an undergraduate degree, while at College B, they take an average of 4.2 years. The pooled standard deviation for this data is 1.6 years

55. Calculate Cohen's d and interpret it.

56. Suppose the mean time to earn an undergraduate degree at College A was 5.2 years. Calculate the effect size and interpret it.

57. You conduct an independent-samples t-test with sample size ten in each of two groups. If you are conducting a two-tailed hypothesis test with $\alpha = 0.01$, what p-values will cause you to reject the null hypothesis?

58. You conduct an independent samples *t*-test with sample size 15 in each group, with the following hypotheses:
H_0: $\mu \geq 110$
H_a: $\mu < 110$
If $\alpha = 0.05$, what *t*-values will cause you to reject the null hypothesis?

10.2: Comparing Two Independent Population Means with Known Population Standard Deviations

Use the following information to answer the next six exercises. College students in the sciences often complain that they must spend more on textbooks each semester than students in the humanities. To test this, you draw random samples of 50 science and 50 humanities students from your college, and record how much each spent last semester on textbooks. Consider the science students to be group one, and the humanities students to be group two.

59. What is the random variable for this study?

60. What are the null and alternative hypotheses for this study?

61. If the 50 science students spent an average of $530 with a sample standard deviation of $20 and the 50 humanities students spent an average of $380 with a sample standard deviation of $15, would you not reject or reject the null hypothesis? Use an alpha level of 0.05. What is your conclusion?

62. What would be your decision, if you were using $\alpha = 0.01$?

10.3: Comparing Two Independent Population Proportions

Use the information to answer the next six exercises. You want to know if proportion of homes with cable television service differs between Community A and Community B. To test this, you draw a random sample of 100 for each and record whether they have cable service.

63. What are the null and alternative hypotheses for this study

64. If 65 households in Community A have cable service, and 78 households in community B, what is the pooled proportion?

65. At $\alpha = 0.03$, will you reject the null hypothesis? What is your conclusion? 65 households in Community A have cable service, and 78 households in community B. 100 households in each community were surveyed.

66. Using an alpha value of 0.01, would you reject the null hypothesis? What is your conclusion? 65 households in Community A have cable service, and 78 households in community B. 100 households in each community were surveyed.

10.4: Matched or Paired Samples

Use the following information to answer the next five exercises. You are interested in whether a particular exercise program helps people lose weight. You conduct a study in which you weigh the participants at the start of the study, and again at the conclusion, after they have participated in the exercise program for six months. You compare the results using a matched-pairs t-test, in which the data is {weight at conclusion – weight at start}. You believe that, on average, the participants will have lost weight after six months on the exercise program.

67. What are the null and alternative hypotheses for this study?

68. Calculate the test statistic, assuming that $\bar{x}_d = -5$, $s_d = 6$, and $n = 30$ (pairs).

69. What are the degrees of freedom for this statistic?

70. Using $\alpha = 0.05$, what is your decision regarding the effectiveness of this program in causing weight loss? What is the conclusion?

71. What would it mean if the *t*-statistic had been 4.56, and what would have been your decision in that case?

11.1: Facts About the Chi-Square Distribution

72. What is the mean and standard deviation for a chi-square distribution with 20 degrees of freedom?

11.2: Goodness-of-Fit Test

Use the following information to answer the next four exercises. Nationally, about 66 percent of high school graduates enroll in higher education. You perform a chi-square goodness of fit test to see if this same proportion applies to your high school's most recent graduating class of 200. Your null hypothesis is that the national distribution also applies to your high school.

73. What are the expected numbers of students from your high school graduating class enrolled and not enrolled in higher education?

74. Fill out the rest of this table.

	Observed (*O*)	**Expected (*E*)**	***O* – *E***	**(*O* – *E*)2**	$\frac{(O-E)^2}{z}$
Enrolled	145				
Not enrolled	55				

Table B10

75. What are the degrees of freedom for this chi-square test?

76. What is the chi-square test statistic and the *p*-value. At the 5% significance level, what do you conclude?

77. For a chi-square distribution with 92 degrees of freedom, the curve _____.

78. For a chi-square distribution with five degrees of freedom, the curve is _____.

11.3: Test of Independence

Use the following information to answer the next four exercises. You are considering conducting a chi-square test of independence for the data in this table, which displays data about cell phone ownership for freshman and seniors at a high school. Your null hypothesis is that cell phone ownership is independent of class standing.

79. Compute the expected values for the cells.

	Cell = Yes	**Cell = No**
Freshman	100	150
Senior	200	50

Table B11

80. Compute $\frac{(O-E)^2}{z}$ for each cell, where O = observed and E = expected.

81. What is the chi-square statistic and degrees of freedom for this study?

82. At the $\alpha = 0.5$ significance level, what is your decision regarding the null hypothesis?

11.4: Test of Homogeneity

83. You conduct a chi-square test of homogeneity for data in a five by two table. What is the degrees of freedom for this test?

11.5: Comparison Summary of the Chi-Square Tests: Goodness-of-Fit, Independence and Homogeneity

84. A 2013 poll in the State of California surveyed people about taxing sugar-sweetened beverages. The results are presented in the following table, and are classified by ethnic group and response type. Are the poll responses independent of the participants' ethnic group? Conduct a hypothesis test at the 5% significance level.

Ethnic Group \ Response Type	Favor	Oppose	No Opinion	Row Total
White / Non-Hispanic	234	433	43	710
Latino	147	106	19	272
African American	24	41	6	71
Asian American	54	48	16	118
Column Total	459	628	84	1171

Table B12

85. In a test of homogeneity, what must be true about the expected value of each cell?

86. Stated in general terms, what are the null and alternative hypotheses for the chi-square test of independence?

87. Stated in general terms, what are the null and alternative hypotheses for the chi-square test of homogeneity?

11.6: Test of a Single Variance

88. A lab test claims to have a variance of no more than five. You believe the variance is greater. What are the null and alternative hypothesis to test this?

Practice Test 3 Solutions

8.1: Confidence Interval, Single Population Mean, Population Standard Deviation Known, Normal

1. $\frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{30}} = 0.73$

2. normal

3. 0.025 or 2.5%; A 95% confidence interval contains 95% of the probability, and excludes five percent, and the five percent excluded is split evenly between the upper and lower tails of the distribution.

4. *z*-score = 1.96; $EBM = z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right) = (1.96)(0.73) = 1.4308$

5. 41 ± 1.43 = (39.57, 42.43); Using the calculator function Zinterval, answer is (40.74, 41.26. Answers differ due to rounding.

6. The *z*-value for a 90% confidence interval is 1.645, so *EBM* = 1.645(0.73) = 1.20085.
The 90% confidence interval is 41 ± 1.20 = (39.80, 42.20).
The calculator function Zinterval answer is (40.78, 41.23). Answers differ due to rounding.

7. The standard error of measurement is: $\frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{50}} = 0.57$

$EBM = z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right) = (1.96)(0.57) = 1.12$

The 95% confidence interval is 41 ± 1.12 = (39.88, 42.12).
The calculator function Zinterval answer is (40.84, 41.16). Answers differ due to rounding.

8. The 99% confidence interval, because it includes all but one percent of the distribution. The 95% confidence interval will be narrower, because it excludes five percent of the distribution.

8.2: Confidence Interval, Single Population Mean, Standard Deviation Unknown, Student's t

9. The *t*-distribution will have more probability in its tails ("thicker tails") and less probability near the mean of the distribution ("shorter in the center").

10. Both distributions are symmetrical and centered at zero.

11. $df = n - 1 = 20 - 1 = 19$

12. You can get the *t*-value from a probability table or a calculator. In this case, for a *t*-distribution with 19 degrees of freedom, and a 95% two-sided confidence interval, the value is 2.093, i.e., $t_{\frac{\alpha}{2}} = 2.093$. The calculator function is invT(0.975, 19).

13. $EBM = t_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right) = (2.093)\left(\frac{0.3}{\sqrt{20}}\right) = 0.140$

98.4 ± 0.14 = (98.26, 98.54).
The calculator function Tinterval answer is (98.26, 98.54).

14. $t_{\frac{\alpha}{2}} = 2.861$. The calculator function is invT(0.995, 19).

$$EBM = t_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right) = (2.861)\left(\frac{0.3}{\sqrt{20}}\right) = 0.192$$

98.4 ± 0.19 = (98.21, 98.59). The calculator function Tinterval answer is (98.21, 98.59).

15. $df = n - 1 = 30 - 1 = 29$. $t_{\frac{\alpha}{2}} = 2.045$

$$EBM = z_t\left(\frac{s}{\sqrt{n}}\right) = (2.045)\left(\frac{0.3}{\sqrt{30}}\right) = 0.112$$

98.4 ± 0.11 = (98.29, 98.51). The calculator function Tinterval answer is (98.29, 98.51).

8.3: Confidence Interval for a Population Proportion

16. $p' = \frac{280}{500} = 0.56$

$q' = 1 - p' = 1 - 0.56 = 0.44$

$s = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.56(0.44)}{500}} = 0.0222$

17. Because you are using the normal approximation to the binomial, $z_{\frac{\alpha}{2}} = 1.96$.

Calculate the error bound for the population (*EBP*):

$$EBP = z_{\frac{\alpha}{2}}\sqrt{\frac{pq}{n}} = 1.96(0.222) = 0.0435$$

Calculate the 95% confidence interval:
0.56 ± 0.0435 = (0.5165, 0.6035).
The calculator function 1-PropZint answer is (0.5165, 0.6035).

18. $z_{\frac{\alpha}{2}} = 1.64$

$$EBP = z_{\frac{\alpha}{2}}\sqrt{\frac{pq}{n}} = 1.64(0.0222) = 0.0364$$

0.56 ± 0.03 = (0.5236, 0.5964). The calculator function 1-PropZint answer is (0.5235, 0.5965)

19. $z_{\frac{\alpha}{2}} = 2.58$

$$EBP = z_{\frac{\alpha}{2}}\sqrt{\frac{pq}{n}} = 2.58(0.0222) = 0.0573$$

0.56 ± 0.05 = (0.5127, 0.6173).
The calculator function 1-PropZint answer is (0.5028, 0.6172).

20. *EBP* = 0.04 (because 4% = 0.04)
$z_{\frac{\alpha}{2}} = 1.96$ for a 95% confidence interval

$$n = \frac{z^2 pq}{EBP^2} = \frac{1.96^2(0.5)(0.5)}{0.04^2} = \frac{0.9604}{0.0016} = 600.25$$

You need 601 subjects (rounding upward from 600.25).

21. $n = \frac{n^2 pq}{EBP^2} = \frac{1.96^2(0.6)(0.4)}{0.04^2} = \frac{0.9220}{0.0016} = 576.24$

You need 577 subjects (rounding upward from 576.24).

22. $n = \frac{n^2 pq}{EBP^2} = \frac{1.96^2(0.5)(0.5)}{0.03^2} = \frac{0.9604}{0.0009} = 1067.11$

You need 1,068 subjects (rounding upward from 1,067.11).

9.1: Null and Alternate Hypotheses

23. H_0: $p = 0.58$
H_a: $p \neq 0.58$

24. H_0: $p \geq 0.58$
H_a: $p < 0.58$

25. H_0: $\mu \geq \$268{,}000$
H_a: $\mu < \$268{,}000$

26. H_a: $\mu \neq 107$

27. H_a: $p \geq 0.25$

9.2: Outcomes and the Type I and Type II Errors

28. a Type I error

29. a Type II error

30. Power = $1 - \beta = 1 - P$(Type II error).

31. The null hypothesis is that the patient does not have cancer. A Type I error would be detecting cancer when it is not present. A Type II error would be not detecting cancer when it is present. A Type II error is more serious, because failure to detect cancer could keep a patient from receiving appropriate treatment.

32. The screening test has a ten percent probability of a Type I error, meaning that ten percent of the time, it will detect TB when it is not present.

33. The screening test has a 20 percent probability of a Type II error, meaning that 20 percent of the time, it will fail to detect TB when it is in fact present.

34. Eighty percent of the time, the screening test will detect TB when it is actually present.

9.3: Distribution Needed for Hypothesis Testing

35. The Student's *t*-test.

36. The normal distribution or *z*-test.

37. The normal distribution with $\mu = p$ and $\sigma = \sqrt{\frac{pq}{n}}$

38. t_{24}. You use the *t*-distribution because you don't know the population standard deviation, and the degrees of freedom are 24 because $df = n - 1$.

39. $\bar{X} \sim N\left(0.95, \frac{0.051}{\sqrt{100}}\right)$

Because you know the population standard deviation, and have a large sample, you can use the normal distribution.

9.4: Rare Events, the Sample, Decision, and Conclusion

40. Fail to reject the null hypothesis, because $\alpha \leq p$

41. Reject the null hypothesis, because $\alpha \geq p$.

42. H_0: $\mu \geq 29.0$"
H_a: $\mu < 29.0$"

43. t_{19}. Because you do not know the population standard deviation, use the *t*-distribution. The degrees of freedom are 19, because $df = n - 1$.

44. The test statistic is –4.4721 and the *p*-value is 0.00013 using the calculator function TTEST.

45. With $\alpha = 0.05$, reject the null hypothesis.

46. With $\alpha = 0.05$, the *p*-value is almost zero using the calculator function TTEST so reject the null hypothesis.

9.5: Additional Information and Full Hypothesis Test Examples

47. The level of significance is five percent.

48. two-tailed

49. one-tailed

50. H_0: $p = 0.8$
H_a: $p \neq 0.8$

51. You will use the normal test for a single population proportion because *np* and *nq* are both greater than five.

10.1: Comparing Two Independent Population Means with Unknown Population Standard Deviations

52. They are matched (paired), because you interviewed married couples.

53. They are independent, because participants were assigned at random to the groups.

54. They are matched (paired), because you collected data twice from each individual.

55. $d = \frac{\bar{x}_1 - \bar{x}_2}{s_{pooled}} = \frac{4.8 - 4.2}{1.6} = 0.375$

This is a small effect size, because 0.375 falls between Cohen's small (0.2) and medium (0.5) effect sizes.

56. $d = \frac{\bar{x}_1 - \bar{x}_2}{s_{pooled}} = \frac{5.2 - 4.2}{1.6} = 0.625$

The effect size is 0.625. By Cohen's standard, this is a medium effect size, because it falls between the medium (0.5) and large (0.8) effect sizes.

57. *p*-value < 0.01.

58. You will only reject the null hypothesis if you get a value significantly below the hypothesized mean of 110.

10.2: Comparing Two Independent Population Means with Known Population Standard Deviations

59. $\bar{X}_1 - \bar{X}_2$, i.e., the mean difference in amount spent on textbooks for the two groups.

60. H_0: $\bar{X}_1 - \bar{X}_2 \leq 0$
H_a: $\bar{X}_1 - \bar{X}_2 > 0$
This could also be written as:
H_0: $\bar{X}_1 \leq \bar{X}_2$
H_a: $\bar{X}_1 > \bar{X}_2$

61. Using the calculator function 2-SampTtest, reject the null hypothesis. At the 5% significance level, there is sufficient evidence to conclude that the science students spend more on textbooks than the humanities students.

62. Using the calculator function 2-SampTtest, reject the null hypothesis. At the 1% significance level, there is sufficient evidence to conclude that the science students spend more on textbooks than the humanities students.

10.3: Comparing Two Independent Population Proportions

63. H_0: $p_A = p_B$
H_a: $p_A \neq p_B$

64. $p_c = \frac{x_A + x_A}{n_A + n_A} = \frac{65 + 78}{100 + 100} = 0.715$

65. Using the calculator function 2-PropZTest, the p-value = 0.0417. Reject the null hypothesis. At the 3% significance level, here is sufficient evidence to conclude that there is a difference between the proportions of households in the two communities that have cable service.

66. Using the calculator function 2-PropZTest, the *p*-value = 0.0417. Do not reject the null hypothesis. At the 1% significance level, there is insufficient evidence to conclude that there is a difference between the proportions of households in the two communities that have cable service.

10.4: Matched or Paired Samples

67. H_0: $\bar{x}_d \geq 0$

H_a: $\bar{x}_d < 0$

68. $t = -4.5644$

69. $df = 30 - 1 = 29$.

70. Using the calculator function TTEST, the *p*-value = 0.00004 so reject the null hypothesis. At the 5% level, there is sufficient evidence to conclude that the participants lost weight, on average.

71. A positive *t*-statistic would mean that participants, on average, gained weight over the six months.

11.1: Facts About the Chi-Square Distribution

72. $\mu = df = 20$

$\sigma = \sqrt{2(df)} = \sqrt{40} = 6.32$

11.2: Goodness-of-Fit Test

73. Enrolled = 200(0.66) = 132. Not enrolled = 200(0.34) = 68

74.

	Observed (O)	Expected (E)	O – E	(O – E)2	$\frac{(O-E)^2}{z}$
Enrolled	145	132	145 – 132 = 13	169	$\frac{169}{132} = 1.280$
Not enrolled	55	68	55 – 68 = –13	169	$\frac{169}{68} = 2.485$

Table B13

75. $df = n - 1 = 2 - 1 = 1$.

76. Using the calculator function Chi-square GOF – Test (in STAT TESTS), the test statistic is 3.7656 and the p-value is 0.0523. Do not reject the null hypothesis. At the 5% significance level, there is insufficient evidence to conclude that high school most recent graduating class distribution of enrolled and not enrolled does not fit that of the national distribution.

77. approximates the normal

78. skewed right

11.3: Test of Independence

79.

	Cell = Yes	Cell = No	Total
Freshman	$\frac{250(300)}{500} = 150$	$\frac{250(200)}{500} = 100$	250
Senior	$\frac{250(300)}{500} = 150$	$\frac{250(200)}{500} = 100$	250
Total	300	200	500

Table B14

80. $\frac{(100-150)^2}{150} = 16.67$

$\frac{(150-100)^2}{100} = 25$

$\frac{(200-100)^2}{150} = 16.67$

$\frac{(50-100)^2}{100} = 25$

81. Chi-square = 16.67 + 25 + 16.67 + 25 = 83.34.
$df = (r - 1)(c - 1) = 1$

82. p-value = P(Chi-square, 83.34) = 0
Reject the null hypothesis.
You could also use the calculator function STAT TESTS Chi-Square – Test.

11.4: Test of Homogeneity

83. The table has five rows and two columns. $df = (r - 1)(c - 1) = (4)(1) = 4$.

11.5: Comparison Summary of the Chi-Square Tests: Goodness-of-Fit, Independence and Homogeneity

84. Using the calculator function (STAT TESTS) Chi-square Test, the p-value = 0. Reject the null hypothesis. At the 5% significance level, there is sufficient evidence to conclude that the poll responses independent of the participants' ethnic group.

85. The expected value of each cell must be at least five.

86. H_0: The variables are independent.
H_a: The variables are not independent.

87. H_0: The populations have the same distribution.
H_a: The populations do not have the same distribution.

11.6: Test of a Single Variance

88. H_0: $\sigma^2 \leq 5$
H_a: $\sigma^2 > 5$

Practice Test 4

12.1 Linear Equations

1. Which of the following equations is/are linear?

a. $y = -3x$
b. $y = 0.2 + 0.74x$
c. $y = -9.4 - 2x$
d. A and B
e. A, B, and C

2. To complete a painting job requires four hours setup time plus one hour per 1,000 square feet. How would you express this information in a linear equation?

3. A statistics instructor is paid a per-class fee of $2,000 plus $100 for each student in the class. How would you express this information in a linear equation?

4. A tutoring school requires students to pay a one-time enrollment fee of $500 plus tuition of $3,000 per year. Express this information in an equation.

12.2: Slope and *Y*-intercept of a Linear Equation

Use the following information to answer the next four exercises. For the labor costs of doing repairs, an auto mechanic charges a flat fee of $75 per car, plus an hourly rate of $55.

5. What are the independent and dependent variables for this situation?

6. Write the equation and identify the slope and intercept.

7. What is the labor charge for a job that takes 3.5 hours to complete?

8. One job takes 2.4 hours to complete, while another takes 6.3 hours. What is the difference in labor costs for these two jobs?

12.3: Scatter Plots

9. Describe the pattern in this scatter plot, and decide whether the *X* and *Y* variables would be good candidates for linear regression.

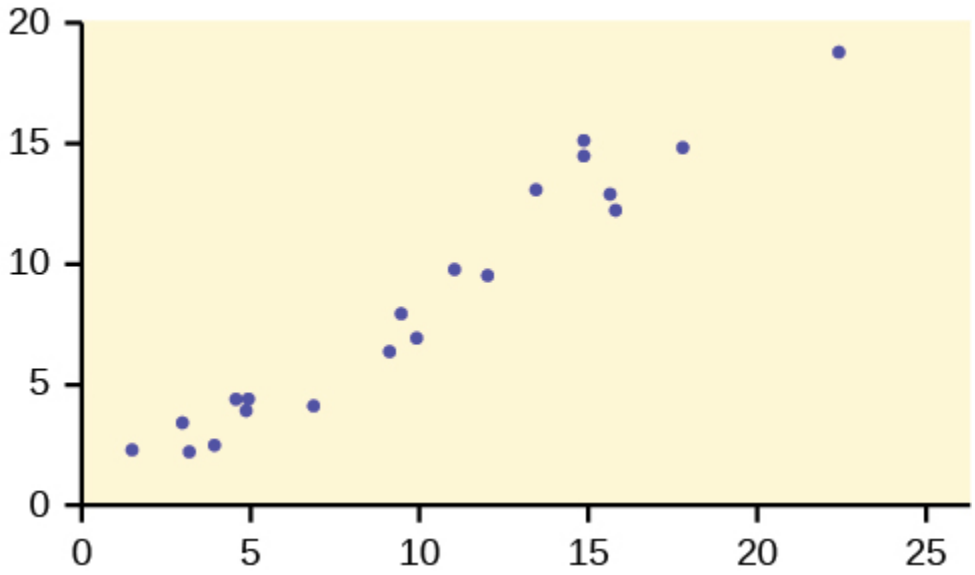

Figure B4

10. Describe the pattern in this scatter plot, and decide whether the *X* and *Y* variables would be good candidates for linear regression.

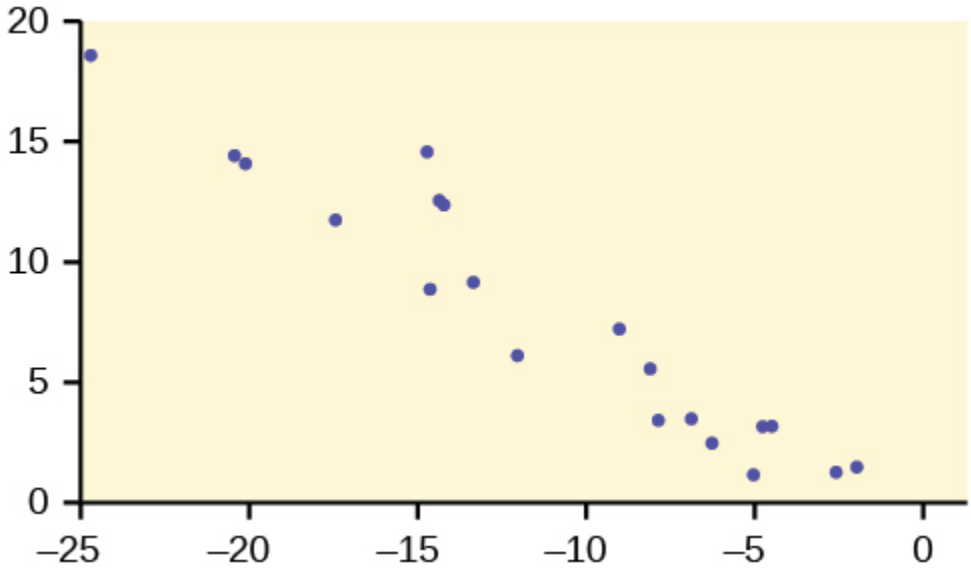

Figure B5

11. Describe the pattern in this scatter plot, and decide whether the *X* and *Y* variables would be good candidates for linear regression.

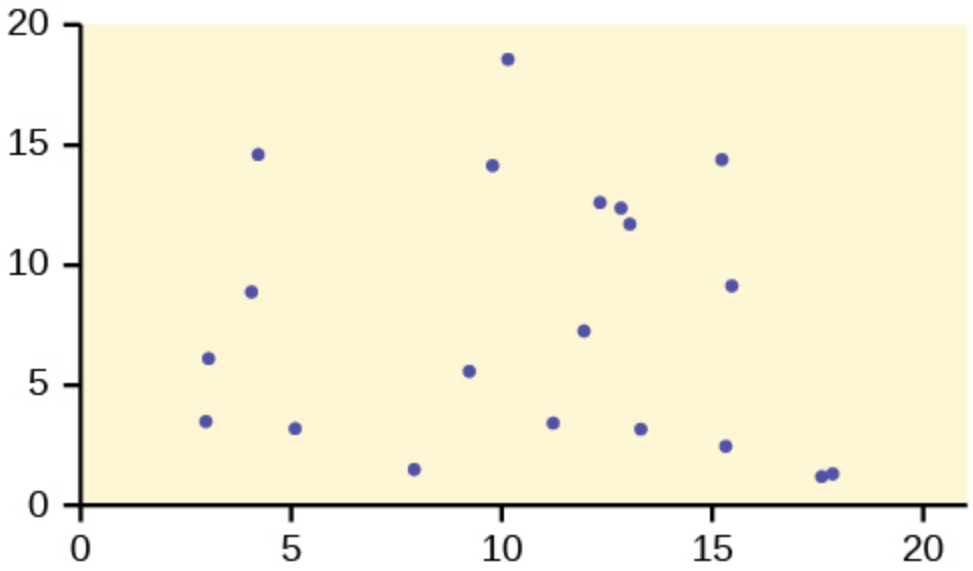

Figure B6

12. Describe the pattern in this scatter plot, and decide whether the X and Y variables would be good candidates for linear regression.

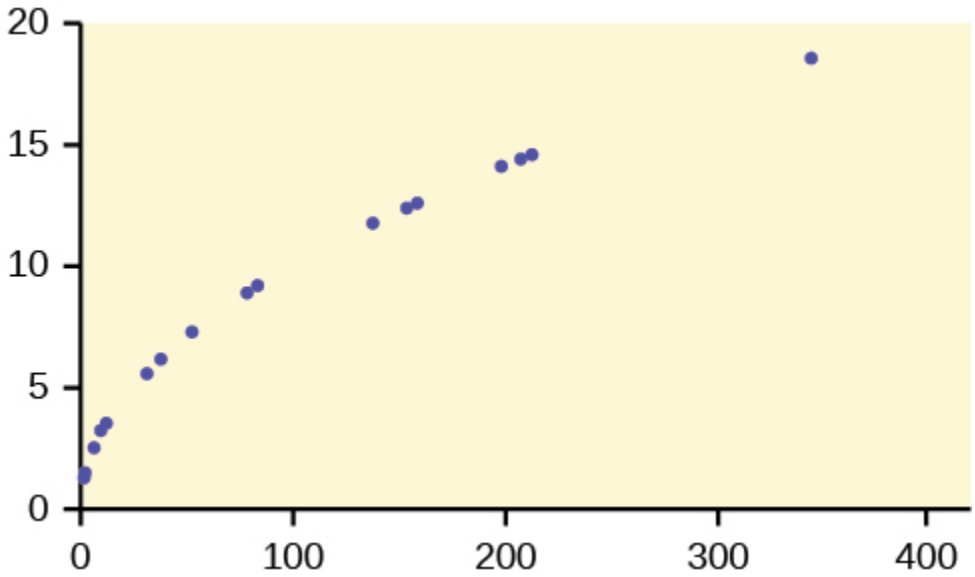

Figure B7

12.4: The Regression Equation

Use the following information to answer the next four exercises. Height (in inches) and weight (In pounds) in a sample of college freshman men have a linear relationship with the following summary statistics:

$\bar{x}$ = 68.4

$\bar{y}$ =141.6

s_x = 4.0

s_y = 9.6

r = 0.73

Let Y = weight and X = height, and write the regression equation in the form:

$\hat{y} = a + bx$

13. What is the value of the slope?

14. What is the value of the *y* intercept?

15. Write the regression equation predicting weight from height in this data set, and calculate the predicted weight for someone 68 inches tall.

12.5: Correlation Coefficient and Coefficient of Determination

16. The correlation between body weight and fuel efficiency (measured as miles per gallon) for a sample of 2,012 model cars is –0.56. Calculate the coefficient of determination for this data and explain what it means.

17. The correlation between high school GPA and freshman college GPA for a sample of 200 university students is 0.32. How much variation in freshman college GPA is not explained by high school GPA?

18. Rounded to two decimal places what correlation between two variables is necessary to have a coefficient of determination of at least 0.50?

12.6: Testing the Significance of the Correlation Coefficient

19. Write the null and alternative hypotheses for a study to determine if two variables are significantly correlated.

20. In a sample of 30 cases, two variables have a correlation of 0.33. Do a *t*-test to see if this result is significant at the α = 0.05 level. Use the formula:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

21. In a sample of 25 cases, two variables have a correlation of 0.45. Do a *t*-test to see if this result is significant at the α = 0.05 level. Use the formula:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

12.7: Prediction

Use the following information to answer the next two exercises. A study relating the grams of potassium (*Y*) to the grams of fiber (*X*) per serving in enriched flour products (bread, rolls, etc.) produced the equation:

$$\hat{y} = 25 + 16x$$

22. For a product with five grams of fiber per serving, what are the expected grams of potassium per serving?

23. Comparing two products, one with three grams of fiber per serving and one with six grams of fiber per serving, what is the expected difference in grams of potassium per serving?

12.8: Outliers

24. In the context of regression analysis, what is the definition of an outlier, and what is a rule of thumb to evaluate if a given value in a data set is an outlier?

25. In the context of regression analysis, what is the definition of an influential point, and how does an influential point differ from an outlier?

26. The least squares regression line for a data set is $\hat{y} = 5 + 0.3x$ and the standard deviation of the residuals is 0.4. Does a case with the values $x = 2$, $y = 6.2$ qualify as an outlier?

27. The least squares regression line for a data set is $\hat{y} = 2.3 - 0.1x$ and the standard deviation of the residuals is 0.13. Does a case with the values $x = 4.1$, $y = 2.34$ qualify as an outlier?

13.1: One-Way ANOVA

28. What are the five basic assumptions to be met if you want to do a one-way ANOVA?

29. You are conducting a one-way ANOVA comparing the effectiveness of four drugs in lowering blood pressure in hypertensive patients. What are the null and alternative hypotheses for this study?

30. What is the primary difference between the independent samples *t*-test and one-way ANOVA?

31. You are comparing the results of three methods of teaching geometry to high school students. The final exam scores X_1, X_2, X_3, for the samples taught by the different methods have the following distributions:
$X_1 \sim N(85, 3.6)$
$X_1 \sim N(82, 4.8)$
$X_1 \sim N(79, 2.9)$

Each sample includes 100 students, and the final exam scores have a range of 0–100. Assuming the samples are independent and randomly selected, have the requirements for conducting a one-way ANOVA been met? Explain why or why not for each assumption.

32. You conduct a study comparing the effectiveness of four types of fertilizer to increase crop yield on wheat farms. When examining the sample results, you find that two of the samples have an approximately normal distribution, and two have an approximately uniform distribution. Is this a violation of the assumptions for conducting a one-way ANOVA?

13.2: The *F* Distribution

Use the following information to answer the next seven exercises. You are conducting a study of three types of feed supplements for cattle to test their effectiveness in producing weight gain among calves whose feed includes one of the supplements. You have four groups of 30 calves (one is a control group receiving the usual feed, but no supplement). You will conduct a one-way ANOVA after one year to see if there are difference in the mean weight for the four groups.

33. What is SS_{within} in this experiment, and what does it mean?

34. What is $SS_{between}$ in this experiment, and what does it mean?

35. What are k and i for this experiment?

36. If $SS_{within} = 374.5$ and $SS_{total} = 621.4$ for this data, what is $SS_{between}$?

37. What are $MS_{between}$, and MS_{within}, for this experiment?

38. What is the F Statistic for this data?

39. If there had been 35 calves in each group, instead of 30, with the sums of squares remaining the same, would the F Statistic be larger or smaller?

13.3: Facts About the *F* Distribution

40. Which of the following numbers are possible F Statistics?

a. 2.47

b. 5.95

c. –3.61

d. 7.28

e. 0.97

41. Histograms $F1$ and $F2$ below display the distribution of cases from samples from two populations, one distributed $F_{3,15}$ and one distributed $F_{5,500}$. Which sample came from which population?

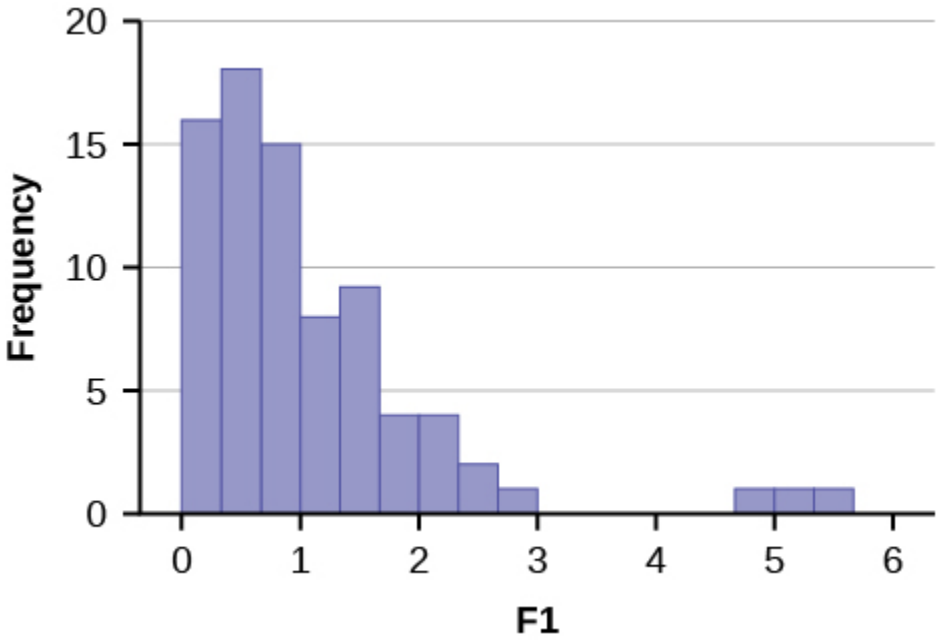

Figure B8

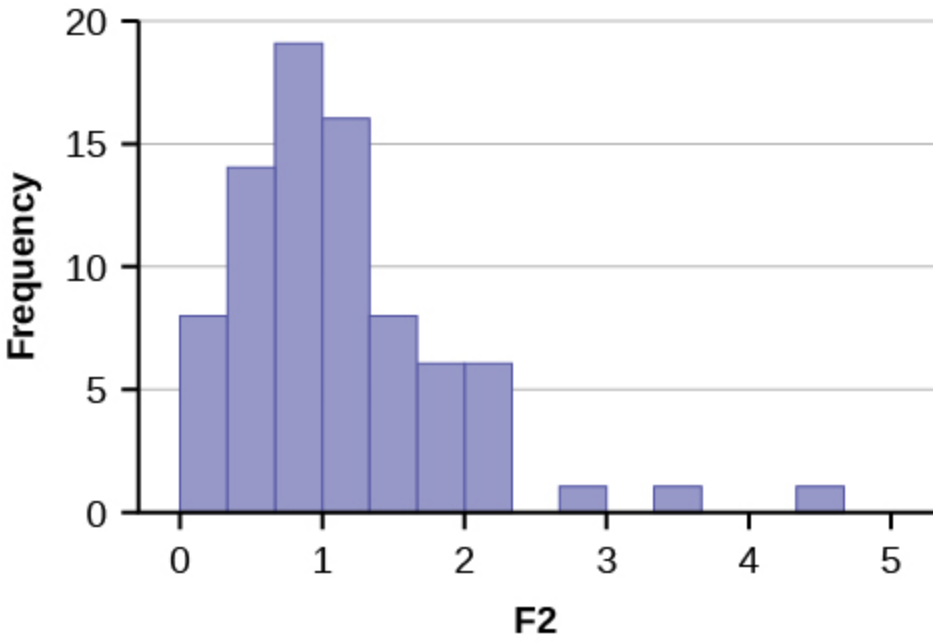

Figure B9

42. The F Statistic from an experiment with $k = 3$ and $n = 50$ is 3.67. At $\alpha = 0.05$, will you reject the null hypothesis?

43. The F Statistic from an experiment with $k = 4$ and $n = 100$ is 4.72. At $\alpha = 0.01$, will you reject the null hypothesis?

13.4: Test of Two Variances

44. What assumptions must be met to perform the F test of two variances?

45. You believe there is greater variance in grades given by the math department at your university than in the English department. You collect all the grades for undergraduate classes in the two departments for a semester, and compute the variance of each, and conduct an F test of two variances. What are the null and alternative hypotheses for this study?

Practice Test 4 Solutions

12.1 Linear Equations

1. e. A, B, and C.
All three are linear equations of the form $y = mx + b$.

2. Let y = the total number of hours required, and x the square footage, measured in units of 1,000. The equation is: $y = x + 4$

3. Let y = the total payment, and x the number of students in a class. The equation is: $y = 100(x) + 2{,}000$

4. Let y = the total cost of attendance, and x the number of years enrolled. The equation is: $y = 3{,}000(x) + 500$

12.2: Slope and Y-intercept of a Linear Equation

5. The independent variable is the hours worked on a car. The dependent variable is the total labor charges to fix a car.

6. Let y = the total charge, and x the number of hours required. The equation is: $y = 55x + 75$
The slope is 55 and the intercept is 75.

7. $y = 55(3.5) + 75 = 267.50$

8. Because the intercept is included in both equations, while you are only interested in the difference in costs, you do not need to include the intercept in the solution. The difference in number of hours required is: 6.3 – 2.4 = 3.9.
Multiply this difference by the cost per hour: 55(3.9) = 214.5.
The difference in cost between the two jobs is $214.50.

12.3: Scatter Plots

9. The X and Y variables have a strong linear relationship. These variables would be good candidates for analysis with linear regression.

10. The *X* and *Y* variables have a strong negative linear relationship. These variables would be good candidates for analysis with linear regression.

11. There is no clear linear relationship between the *X* and *Y* variables, so they are not good candidates for linear regression.

12. The *X* and *Y* variables have a strong positive relationship, but it is curvilinear rather than linear. These variables are not good candidates for linear regression.

12.4: The Regression Equation

13. $r\left(\frac{s_y}{s_x}\right) = 0.73\left(\frac{9.6}{4.0}\right) = 1.752 \approx 1.75$

14. $a = \bar{y} - b\bar{x} = 141.6 - 1.752(68.4) = 21.7632 \approx 21.76$

15. $\hat{y} = 21.76 + 1.75(68) = 140.76$

12.5: Correlation Coefficient and Coefficient of Determination

16. The coefficient of determination is the square of the correlation, or r^2.

For this data, $r^2 = (-0.56)2 = 0.3136 \approx 0.31$ or 31%. This means that 31 percent of the variation in fuel efficiency can be explained by the bodyweight of the automobile.

17. The coefficient of determination = $0.32^2 = 0.1024$. This is the amount of variation in freshman college GPA that can be explained by high school GPA. The amount that cannot be explained is $1 - 0.1024 = 0.8976 \approx 0.90$. So about 90 percent of variance in freshman college GPA in this data is not explained by high school GPA.

18. $r = \sqrt{r^2}$

$\sqrt{0.5} = 0.707106781 \approx 0.71$

You need a correlation of 0.71 or higher to have a coefficient of determination of at least 0.5.

12.6: Testing the Significance of the Correlation Coefficient

19. H_0: $\rho = 0$

H_a: $\rho \neq 0$

20. $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.33\sqrt{30-2}}{\sqrt{1-0.33^2}} = 1.85$

The critical value for $\alpha = 0.05$ for a two-tailed test using the t_{29} distribution is 2.045. Your value is less than this, so you fail to reject the null hypothesis and conclude that the study produced no evidence that the variables are significantly correlated. Using the calculator function tcdf, the *p*-value is 2tcdf(1.85, 10^99, 29) = 0.0373. Do not reject the null hypothesis and conclude that the study produced no evidence that the variables are significantly correlated.

21. $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.45\sqrt{25-2}}{\sqrt{1-0.45^2}} = 2.417$

The critical value for $\alpha = 0.05$ for a two-tailed test using the t_{24} distribution is 2.064. Your value is greater than this, so you reject the null hypothesis and conclude that the study produced evidence that the variables are significantly correlated.
Using the calculator function tcdf, the p-value is 2tcdf(2.417, 10^99, 24) = 0.0118. Reject the null hypothesis and conclude that the study produced evidence that the variables are significantly correlated.

12.7: Prediction

22. $\hat{y} = 25 + 16(5) = 105$

23. Because the intercept appears in both predicted values, you can ignore it in calculating a predicted difference score. The difference in grams of fiber per serving is 6 – 3 = 3 and the predicted difference in grams of potassium per serving is (16)(3) = 48.

12.8: Outliers

24. An outlier is an observed value that is far from the least squares regression line. A rule of thumb is that a point more than two standard deviations of the residuals from its predicted value on the least squares regression line is an outlier.

25. An influential point is an observed value in a data set that is far from other points in the data set, in a horizontal direction. Unlike an outlier, an influential point is determined by its relationship with other values in the data set, not by its relationship to the regression line.

26. The predicted value for y is: $\hat{y} = 5 + 0.3x = 5.6$. The value of 6.2 is less than two standard deviations from the predicted value, so it does not qualify as an outlier.
Residual for (2, 6.2): 6.2 – 5.6 = 0.6 (0.6 < 2(0.4))

27. The predicted value for y is: $\hat{y}$ = 2.3 – 0.1(4.1) = 1.89. The value of 2.32 is more than two standard deviations from the predicted value, so it qualifies as an outlier.
Residual for (4.1, 2.34): 2.32 – 1.89 = 0.43 (0.43 > 2(0.13))

13.1: One-Way ANOVA

28.

1. Each sample is drawn from a normally distributed population
2. All samples are independent and randomly selected.
3. The populations from which the samples are draw have equal standard deviations.
4. The factor is a categorical variable.
5. The response is a numerical variable.

29. H_0: $\mu 1 = \mu 2 = \mu 3 = \mu 4$
H_a: At least two of the group means $\mu 1$, $\mu 2$, $\mu 3$, $\mu 4$ are not equal.

30. The independent samples *t*-test can only compare means from two groups, while one-way ANOVA can compare means of more than two groups.

31. Each sample appears to have been drawn from a normally distributed populations, the factor is a categorical variable (method), the outcome is a numerical variable (test score), and you were told the samples were independent and randomly selected, so those requirements are met. However, each sample has a different standard deviation, and this suggests that the populations from which they were drawn also have different standard deviations, which is a violation of an assumption for one-way ANOVA. Further statistical testing will be necessary to test the assumption of equal variance before proceeding with the analysis.

32. One of the assumptions for a one-way ANOVA is that the samples are drawn from normally distributed populations. Since two of your samples have an approximately uniform distribution, this casts doubt on whether this assumption has been met. Further statistical testing will be necessary to determine if you can proceed with the analysis.

13.2: The *F* Distribution

33. SS_{within} is the sum of squares within groups, representing the variation in outcome that cannot be attributed to the different feed supplements, but due to individual or chance factors among the calves in each group.

34. $SS_{between}$ is the sum of squares between groups, representing the variation in outcome that can be attributed to the different feed supplements.

35. k = the number of groups = 4
n_1 = the number of cases in group 1 = 30
n = the total number of cases = 4(30) = 120

36. $SS_{total} = SS_{within} + SS_{between}$ so $SS_{between} = SS_{total} - SS_{within}$
621.4 – 374.5 = 246.9

37. The mean squares in an ANOVA are found by dividing each sum of squares by its respective degrees of freedom (df).
For SS_{total}, $df = n - 1 = 120 - 1 = 119$.
For $SS_{between}$, df = k – 1 = 4 – 1 = 3.
For SS_{within}, df = 120 – 4 = 116.
$MS_{between} = \frac{246.9}{3} = 82.3$

$MS_{within} = \frac{374.5}{116} = 3.23$

38. $F = \frac{MS_{between}}{MS_{within}} = \frac{82.3}{3.23} = 25.48$

39. It would be larger, because you would be dividing by a smaller number. The value of $MS_{between}$ would not change with a change of sample size, but the value of MS_{within} would be smaller, because you would be dividing by a larger number (df_{within} would be 136, not 116). Dividing a constant by a smaller number produces a larger result.

13.3: Facts About the *F* Distribution

40. All but choice c, –3.61. *F* Statistics are always greater than or equal to 0.

41. As the degrees of freedom increase in an *F* distribution, the distribution becomes more nearly normal. Histogram *F*2 is closer to a normal distribution than histogram *F*1, so the sample displayed in histogram *F*1 was drawn from the $F_{3,15}$ population, and the sample displayed in histogram *F*2 was drawn from the $F_{5,500}$ population.

42. Using the calculator function Fcdf, *p*-value = Fcdf(3.67, 1E, 3,50) = 0.0182. Reject the null hypothesis.

43. Using the calculator function Fcdf, *p*-value = Fcdf(4.72, 1E, 4, 100) = 0.0016 Reject the null hypothesis.

13.4: Test of Two Variances

44. The samples must be drawn from populations that are normally distributed, and must be drawn from independent populations.

45. Let σ_M^2 = variance in math grades, and σ_E^2 = variance in English grades.

H_0: $\sigma_M^2 \leq \sigma_E^2$

H_a: $\sigma_M^2 > \sigma_E^2$

Practice Final Exam 1

Use the following information to answer the next two exercises: An experiment consists of tossing two, 12-sided dice (the numbers 1–12 are printed on the sides of each die).

- Let Event *A* = both dice show an even number.
- Let Event *B* = both dice show a number more than eight

1. Events *A* and *B* are:

a. mutually exclusive.

b. independent.

c. mutually exclusive and independent.

d. neither mutually exclusive nor independent.

2. Find $P(A|B)$.

a. $\frac{2}{4}$

b. $\frac{16}{144}$

c. $\frac{4}{16}$

d. $\frac{2}{144}$

3. Which of the following are TRUE when we perform a hypothesis test on matched or paired samples?

a. Sample sizes are almost never small.

b. Two measurements are drawn from the same pair of individuals or objects.

c. Two sample means are compared to each other.

d. Answer choices b and c are both true.

Use the following information to answer the next two exercises: One hundred eighteen students were asked what type of color their bedrooms were painted: light colors, dark colors, or vibrant colors. The results were tabulated according to gender.

	Light colors	Dark colors	Vibrant colors
Female	20	22	28
Male	10	30	8

Table B15

4. Find the probability that a randomly chosen student is male or has a bedroom painted with light colors.

a. $\frac{10}{118}$

b. $\frac{68}{118}$

c. $\frac{48}{118}$

d. $\frac{10}{48}$

5. Find the probability that a randomly chosen student is male given the student's bedroom is painted with dark colors.

a. $\frac{30}{118}$

b. $\frac{30}{48}$

c. $\frac{22}{118}$

d. $\frac{30}{52}$

Use the following information to answer the next two exercises: We are interested in the number of times a teenager must be reminded to do his or her chores each week. A survey of 40 mothers was conducted. **Table B16** shows the results of the survey.

x	$P(x)$
0	$\frac{2}{40}$
1	$\frac{5}{40}$
2	
3	$\frac{14}{40}$
4	$\frac{7}{40}$
5	$\frac{4}{40}$

Table B16

6. Find the probability that a teenager is reminded two times.

a. 8

b. $\frac{8}{40}$

c. $\frac{6}{40}$

d. 2

7. Find the expected number of times a teenager is reminded to do his or her chores.

a. 15

b. 2.78

c. 1.0

d. 3.13

Use the following information to answer the next two exercises: On any given day, approximately 37.5% of the cars parked in the De Anza parking garage are parked crookedly. We randomly survey 22 cars. We are interested in the number of cars that are parked crookedly.

8. For every 22 cars, how many would you expect to be parked crookedly, on average?

a. 8.25

b. 11

c. 18

d. 7.5

9. What is the probability that at least ten of the 22 cars are parked crookedly.

a. 0.1263

b. 0.1607

c. 0.2870

d. 0.8393

10. Using a sample of 15 Stanford-Binet IQ scores, we wish to conduct a hypothesis test. Our claim is that the mean IQ score on the Stanford-Binet IQ test is more than 100. It is known that the standard deviation of all Stanford-Binet IQ scores is 15 points. The correct distribution to use for the hypothesis test is:

a. Binomial

b. Student's *t*

c. Normal

d. Uniform

Use the following information to answer the next three exercises: De Anza College keeps statistics on the pass rate of students who enroll in math classes. In a sample of 1,795 students enrolled in Math 1A (1st quarter calculus), 1,428 passed the course. In a sample of 856 students enrolled in Math 1B (2nd quarter calculus), 662 passed. In general, are the pass rates of Math 1A and Math 1B statistically the same? Let A = the subscript for Math 1A and B = the subscript for Math 1B.

11. If you were to conduct an appropriate hypothesis test, the alternate hypothesis would be:

a. H_a: $p_A = p_B$

b. H_a: $p_A > p_B$

c. H_o: $p_A = p_B$

d. H_a: $p_A \neq p_B$

12. The Type I error is to:

a. conclude that the pass rate for Math 1A is the same as the pass rate for Math 1B when, in fact, the pass rates are different.

b. conclude that the pass rate for Math 1A is different than the pass rate for Math 1B when, in fact, the pass rates are the same.

c. conclude that the pass rate for Math 1A is greater than the pass rate for Math 1B when, in fact, the pass rate for Math 1A is less than the pass rate for Math 1B.

d. conclude that the pass rate for Math 1A is the same as the pass rate for Math 1B when, in fact, they are the same.

13. The correct decision is to:

a. reject H_0

b. not reject H_0

c. There is not enough information given to conduct the hypothesis test

Kia, Alejandra, and Iris are runners on the track teams at three different schools. Their running times, in minutes, and the statistics for the track teams at their respective schools, for a one mile run, are given in the table below:

	Running Time	School Average Running Time	School Standard Deviation
Kia	4.9	5.2	0.15
Alejandra	4.2	4.6	0.25
Iris	4.5	4.9	0.12

Table B17

14. Which student is the BEST when compared to the other runners at her school?

a. Kia

b. Alejandra

c. Iris

d. Impossible to determine

Use the following information to answer the next two exercises: The following adult ski sweater prices are from the Gorsuch Ltd. Winter catalog: $212, $292, $278, $199, $280, $236

Assume the underlying sweater price population is approximately normal. The null hypothesis is that the mean price of adult ski sweaters from Gorsuch Ltd. is at least $275.

15. The correct distribution to use for the hypothesis test is:

a. Normal

b. Binomial

c. Student's *t*

d. Exponential

16. The hypothesis test:

a. is two-tailed.

b. is left-tailed.

c. is right-tailed.

d. has no tails.

17. Sara, a statistics student, wanted to determine the mean number of books that college professors have in their office. She randomly selected two buildings on campus and asked each professor in the selected buildings how many books are in his or her office. Sara surveyed 25 professors. The type of sampling selected is

a. simple random sampling.

b. systematic sampling.

c. cluster sampling.

d. stratified sampling.

18. A clothing store would use which measure of the center of data when placing orders for the typical "middle" customer?

a. mean

b. median

c. mode

d. IQR

19. In a hypothesis test, the *p*-value is

a. the probability that an outcome of the data will happen purely by chance when the null hypothesis is true.

b. called the preconceived alpha.

c. compared to beta to decide whether to reject or not reject the null hypothesis.

d. Answer choices A and B are both true.

Use the following information to answer the next three exercises: A community college offers classes 6 days a week: Monday through Saturday. Maria conducted a study of the students in her classes to determine how many days per week the students who are in her classes come to campus for classes. In each of her 5 classes she randomly selected 10 students and asked them how many days they come to campus for classes. Each of her classes are the same size. The results of her survey are summarized in **Table B18**.

Number of Days on Campus	Frequency	Relative Frequency	Cumulative Relative Frequency
1	2		
2	12	.24	
3	10	.20	
4			.98
5	0		
6	1	.02	1.00

Table B18

20. Combined with convenience sampling, what other sampling technique did Maria use?

a. simple random

b. systematic

c. cluster

d. stratified

21. How many students come to campus for classes four days a week?

a. 49

b. 25

c. 30

d. 13

22. What is the 60^{th} percentile for the this data?

a. 2

b. 3

c. 4

d. 5

Use the following information to answer the next two exercises: The following data are the results of a random survey of 110 Reservists called to active duty to increase security at California airports.

Number of Dependents	Frequency
0	11
1	27
2	33
3	20
4	19

Table B19

23. Construct a 95% confidence interval for the true population mean number of dependents of Reservists called to active duty to increase security at California airports.

a. (1.85, 2.32)
b. (1.80, 2.36)
c. (1.97, 2.46)
d. (1.92, 2.50)

24. The 95% confidence interval above means:

a. Five percent of confidence intervals constructed this way will not contain the true population aveage number of dependents.
b. We are 95% confident the true population mean number of dependents falls in the interval.
c. Both of the above answer choices are correct.
d. None of the above.

25. $X \sim U(4, 10)$. Find the 30^{th} percentile.

a. 0.3000
b. 3
c. 5.8
d. 6.1

26. If $X \sim Exp(0.8)$, then $P(x < \mu) =$ _____

a. 0.3679
b. 0.4727
c. 0.6321
d. cannot be determined

27. The lifetime of a computer circuit board is normally distributed with a mean of 2,500 hours and a standard deviation of 60 hours. What is the probability that a randomly chosen board will last at most 2,560 hours?

a. 0.8413
b. 0.1587
c. 0.3461
d. 0.6539

28. A survey of 123 reservists called to active duty as a result of the September 11, 2001, attacks was conducted to determine the proportion that were married. Eighty-six reported being married. Construct a 98% confidence interval for the true population proportion of reservists called to active duty that are married.

a. (0.6030, 0.7954)
b. (0.6181, 0.7802)
c. (0.5927, 0.8057)
d. (0.6312, 0.7672)

29. Winning times in 26 mile marathons run by world class runners average 145 minutes with a standard deviation of 14 minutes. A sample of the last ten marathon winning times is collected. Let x = mean winning times for ten marathons. The distribution for x is:

a. $N\left(145, \frac{14}{\sqrt{10}}\right)$
b. $N(145,14)$
c. t_9
d. t_{10}

30. Suppose that Phi Beta Kappa honors the top one percent of college and university seniors. Assume that grade point means (GPA) at a certain college are normally distributed with a 2.5 mean and a standard deviation of 0.5. What would be the minimum GPA needed to become a member of Phi Beta Kappa at that college?

a. 3.99
b. 1.34
c. 3.00
d. 3.66

The number of people living on American farms has declined steadily during the 20th century. Here are data on the farm population (in millions of persons) from 1935 to 1980.

Year	1935	1940	1945	1950	1955	1960	1965	1970	1975	1980
Population	32.1	30.5	24.4	23.0	19.1	15.6	12.4	9.7	8.9	7.2

Table B20

31. The linear regression equation is $\hat{y} = 1166.93 - 0.5868x$. What was the expected farm population (in millions of persons) for 1980?

a. 7.2
b. 5.1
c. 6.0
d. 8.0

32. In linear regression, which is the best possible *SSE*?

a. 13.46
b. 18.22
c. 24.05
d. 16.33

33. In regression analysis, if the correlation coefficient is close to one what can be said about the best fit line?

a. It is a horizontal line. Therefore, we can not use it.
b. There is a strong linear pattern. Therefore, it is most likely a good model to be used.
c. The coefficient correlation is close to the limit. Therefore, it is hard to make a decision.
d. We do not have the equation. Therefore, we cannot say anything about it.

Use the following information to answer the next three exercises: A study of the career plans of young women and men sent questionnaires to all 722 members of the senior class in the College of Business Administration at the University of Illinois. One question asked which major within the business program the student had chosen. Here are the data from the students who responded.

	Female	Male
Accounting	68	56
Administration	91	40
Economics	5	6
Finance	61	59

Table B21 Does the data suggest that there is a relationship between the gender of students and their choice of major?

34. The distribution for the test is:

a. Chi^2_8.

b. Chi^2_3.

c. t_{721}.

d. $N(0, 1)$.

35. The expected number of female who choose finance is:

a. 37.

b. 61.

c. 60.

d. 70.

36. The *p*-value is 0.0127 and the level of significance is 0.05. The conclusion to the test is:

a. there is insufficient evidence to conclude that the choice of major and the gender of the student are not independent of each other.

b. there is sufficient evidence to conclude that the choice of major and the gender of the student are not independent of each other.

c. there is sufficient evidence to conclude that students find economics very hard.

d. there is in sufficient evidence to conclude that more females prefer administration than males.

37. An agency reported that the work force nationwide is composed of 10% professional, 10% clerical, 30% skilled, 15% service, and 35% semiskilled laborers. A random sample of 100 San Jose residents indicated 15 professional, 15 clerical, 40 skilled, 10 service, and 20 semiskilled laborers. At $\alpha = 0.10$ does the work force in San Jose appear to be consistent with the agency report for the nation? Which kind of test is it?

a. Chi^2 goodness of fit

b. Chi^2 test of independence

c. Independent groups proportions

d. Unable to determine

Practice Final Exam 1 Solutions

Solutions

1. b. independent

2. c. $\frac{4}{16}$

3. b. Two measurements are drawn from the same pair of individuals or objects.

4. b. $\frac{68}{118}$

5. d. $\frac{30}{52}$

6. b. $\frac{8}{40}$

7. b. 2.78

8. a. 8.25

9. c. 0.2870

10. c. Normal

11. d. H_a: $p_A \neq p_B$

12. b. conclude that the pass rate for Math 1A is different than the pass rate for Math 1B when, in fact, the pass rates are the same.

13. b. not reject H_0

14. c. Iris

15. c. Student's *t*

16. b. is left-tailed.

17. c. cluster sampling

18. b. median

19. a. the probability that an outcome of the data will happen purely by chance when the null hypothesis is true.

20. d. stratified

21. b. 25

22. c. 4

23. a. (1.85, 2.32)

24. c. Both above are correct.

25. c. 5.8

26. c. 0.6321

27. a. 0.8413

28. a. (0.6030, 0.7954)

29. a. $N\left(145, \frac{14}{\sqrt{10}}\right)$

30. d. 3.66

31. b. 5.1

32. a. 13.46

33. b. There is a strong linear pattern. Therefore, it is most likely a good model to be used.

34. b. Chi^2_3 .

35. d. 70

36. b. There is sufficient evidence to conclude that the choice of major and the gender of the student are not independent of each other.

37. a. Chi^2 goodness-of-fit

Practice Final Exam 2

1. A study was done to determine the proportion of teenagers that own a car. The population proportion of teenagers that own a car is the:

a. statistic.
b. parameter.
c. population.
d. variable.

Use the following information to answer the next two exercises:

value	frequency
0	1
1	4
2	7
3	9
6	4

Table B22

2. The box plot for the data is:

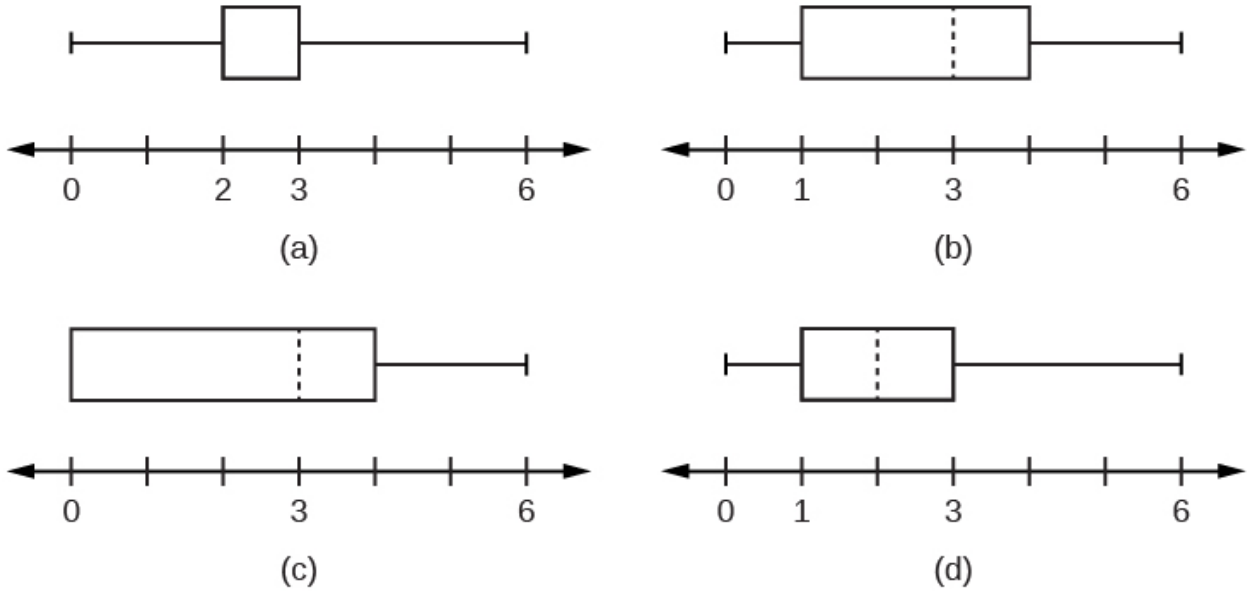

Figure B10

3. If six were added to each value of the data in the table, the 15th percentile of the new list of values is:

a. six
b. one
c. seven
d. eight

Use the following information to answer the next two exercises: Suppose that the probability of a drought in any independent year is 20%. Out of those years in which a drought occurs, the probability of water rationing is ten percent. However, in any year, the probability of water rationing is five percent.

4. What is the probability of both a drought and water rationing occurring?

a. 0.05
b. 0.01
c. 0.02
d. 0.30

5. Which of the following is true?

a. Drought and water rationing are independent events.
b. Drought and water rationing are mutually exclusive events.
c. None of the above

Use the following information to answer the next two exercises: Suppose that a survey yielded the following data:

gender	apple	pumpkin	pecan
female	40	10	30
male	20	30	10

Table B23 Favorite Pie

6. Suppose that one individual is randomly chosen. The probability that the person's favorite pie is apple or the person is male is _____.

a. $\frac{40}{60}$

b. $\frac{60}{140}$

c. $\frac{120}{140}$

d. $\frac{100}{140}$

7. Suppose H_0 is: Favorite pie and gender are independent. The *p*-value is _____.

a. ≈ 0

b. 1

c. 0.05

d. cannot be determined

Use the following information to answer the next two exercises: Let's say that the probability that an adult watches the news at least once per week is 0.60. We randomly survey 14 people. Of interest is the number of people who watch the news at least once per week.

8. Which of the following statements is FALSE?

a. $X \sim B(14\ 0.60)$

b. The values for x are: $\{1, 2, 3, \ldots, 14\}$.

c. $\mu = 8.4$

d. $P(X = 5) = 0.0408$

9. Find the probability that at least six adults watch the news at least once per week.

a. $\frac{6}{14}$

b. 0.8499

c. 0.9417

d. 0.6429

10. The following histogram is most likely to be a result of sampling from which distribution?

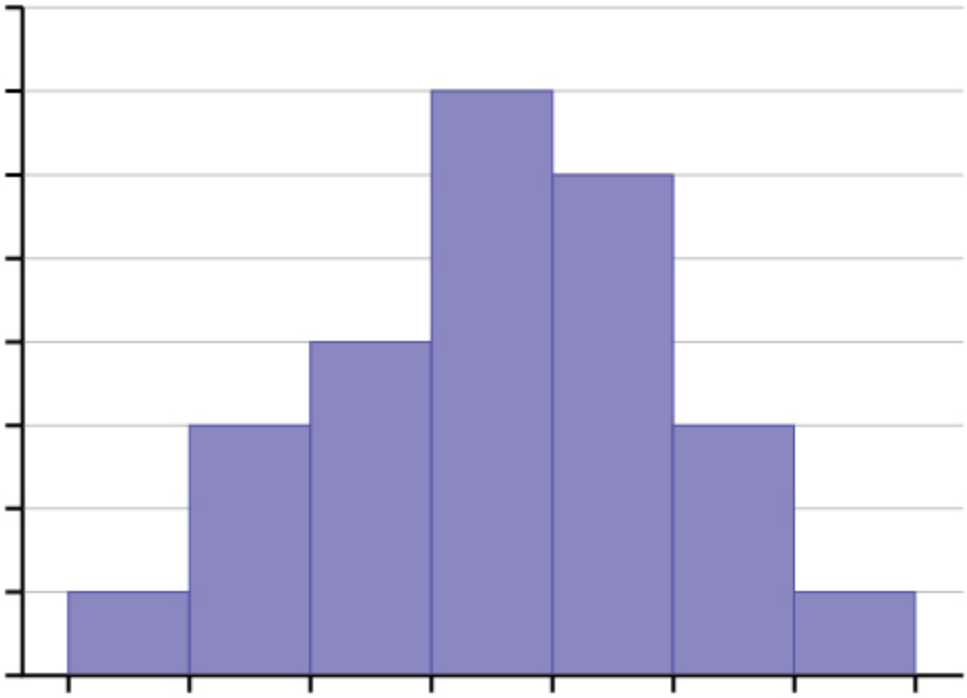

Figure B11

a. chi-square with $df = 6$
b. exponential
c. uniform
d. binomial

11. The ages of campus day and evening students is known to be normally distributed. A sample of six campus day and evening students reported their ages (in years) as: {18, 35, 27, 45, 20, 20}. What is the error bound for the 90% confidence interval of the true average age?

a. 11.2
b. 22.3
c. 17.5
d. 8.7

12. If a normally distributed random variable has $\mu = 0$ and $\sigma = 1$, then 97.5% of the population values lie above:

a. –1.96.
b. 1.96.
c. 1.
d. –1.

Use the following information to answer the next three exercises. The amount of money a customer spends in one trip to the supermarket is known to have an exponential distribution. Suppose the average amount of money a customer spends in one trip to the supermarket is $72.

13. What is the probability that one customer spends less than $72 in one trip to the supermarket?

a. 0.6321
b. 0.5000
c. 0.3714
d. 1

14. How much money altogether would you expect the next five customers to spend in one trip to the supermarket (in dollars)?

a. 72

b. $\frac{72^2}{5}$

c. 5184

d. 360

15. If you want to find the probability that the mean amount of money 50 customers spend in one trip to the supermarket is less than $60, the distribution to use is:

a. $N(72, 72)$

b. $N\left(72, \frac{72}{\sqrt{50}}\right)$

c. $Exp(72)$

d. $Exp\left(\frac{1}{72}\right)$

Use the following information to answer the next three exercises: The amount of time it takes a fourth grader to carry out the trash is uniformly distributed in the interval from one to ten minutes.

16. What is the probability that a randomly chosen fourth grader takes more than seven minutes to take out the trash?

a. $\frac{3}{9}$

b. $\frac{7}{9}$

c. $\frac{3}{10}$

d. $\frac{7}{10}$

17. Which graph best shows the probability that a randomly chosen fourth grader takes more than six minutes to take out the trash given that he or she has already taken more than three minutes?

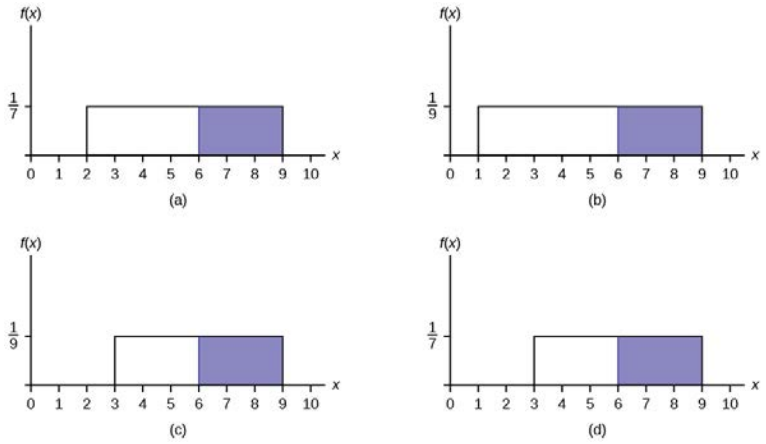

Figure B12

18. We should expect a fourth grader to take how many minutes to take out the trash?

a. 4.5

b. 5.5

c. 5

d. 10

Use the following information to answer the next three exercises: At the beginning of the quarter, the amount of time a

student waits in line at the campus cafeteria is normally distributed with a mean of five minutes and a standard deviation of 1.5 minutes.

19. What is the 90^{th} percentile of waiting times (in minutes)?

a. 1.28
b. 90
c. 7.47
d. 6.92

20. The median waiting time (in minutes) for one student is:

a. 5.
b. 50.
c. 2.5.
d. 1.5.

21. Find the probability that the average wait time for ten students is at most 5.5 minutes.

a. 0.6301
b. 0.8541
c. 0.3694
d. 0.1459

22. A sample of 80 software engineers in Silicon Valley is taken and it is found that 20% of them earn approximately $50,000 per year. A point estimate for the true proportion of engineers in Silicon Valley who earn $50,000 per year is:

a. 16.
b. 0.2.
c. 1.
d. 0.95.

23. If $P(Z < z_\alpha) = 0.1587$ where $Z \sim N(0, 1)$, then α is equal to:

a. –1.
b. 0.1587.
c. 0.8413.
d. 1.

24. A professor tested 35 students to determine their entering skills. At the end of the term, after completing the course, the same test was administered to the same 35 students to study their improvement. This would be a test of:

a. independent groups.
b. two proportions.
c. matched pairs, dependent groups.
d. exclusive groups.

A math exam was given to all the third grade children attending ABC School. Two random samples of scores were taken.

	n	$\bar{x}$	***s***
Boys	55	82	5
Girls	60	86	7

Table B24

25. Which of the following correctly describes the results of a hypothesis test of the claim, “There is a difference between the mean scores obtained by third grade girls and boys at the 5% level of significance”?

a. Do not reject H_0. There is insufficient evidence to conclude that there is a difference in the mean scores.
b. Do not reject H_0. There is sufficient evidence to conclude that there is a difference in the mean scores.

c. Reject H_0. There is insufficient evidence to conclude that there is no difference in the mean scores.

d. Reject H_0. There is sufficient evidence to conclude that there is a difference in the mean scores.

26. In a survey of 80 males, 45 had played an organized sport growing up. Of the 70 females surveyed, 25 had played an organized sport growing up. We are interested in whether the proportion for males is higher than the proportion for females. The correct conclusion is that:

a. there is insufficient information to conclude that the proportion for males is the same as the proportion for females.

b. there is insufficient information to conclude that the proportion for males is not the same as the proportion for females.

c. there is sufficient evidence to conclude that the proportion for males is higher than the proportion for females.

d. not enough information to make a conclusion.

27. From past experience, a statistics teacher has found that the average score on a midterm is 81 with a standard deviation of 5.2. This term, a class of 49 students had a standard deviation of 5 on the midterm. Do the data indicate that we should reject the teacher's claim that the standard deviation is 5.2? Use $\alpha = 0.05$.

a. Yes

b. No

c. Not enough information given to solve the problem

28. Three loading machines are being compared. Ten samples were taken for each machine. Machine I took an average of 31 minutes to load packages with a standard deviation of two minutes. Machine II took an average of 28 minutes to load packages with a standard deviation of 1.5 minutes. Machine III took an average of 29 minutes to load packages with a standard deviation of one minute. Find the *p*-value when testing that the average loading times are the same.

a. *p*-value is close to zero

b. *p*-value is close to one

c. not enough information given to solve the problem

Use the following information to answer the next three exercises: A corporation has offices in different parts of the country. It has gathered the following information concerning the number of bathrooms and the number of employees at seven sites:

Number of employees *x*	650	730	810	900	102	107	1150
Number of bathrooms *y*	40	50	54	61	82	110	121

Table B25

29. Is the correlation between the number of employees and the number of bathrooms significant?

a. Yes

b. No

c. Not enough information to answer question

30. The linear regression equation is:

a. $\hat{y} = 0.0094 - 79.96x$

b. $\hat{y} = 79.96 + 0.0094x$

c. $\hat{y} = 79.96 - 0.0094x$

d. $\hat{y} = -0.0094 + 79.96x$

31. If a site has 1,150 employees, approximately how many bathrooms should it have?

a. 69

b. 91

c. 91,954

d. We should not be estimating here.

32. Suppose that a sample of size ten was collected, with $\bar{x} = 4.4$ and $s = 1.4$. H_0: $\sigma^2 = 1.6$ vs. H_a: $\sigma^2 \neq 1.6$. Which graph best describes the results of the test?

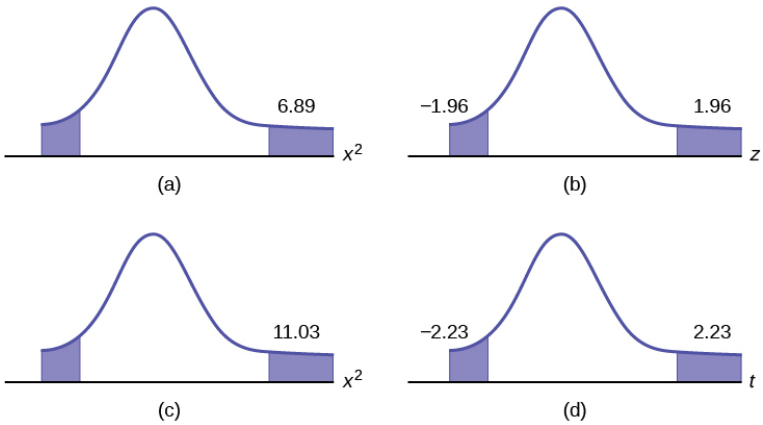

Figure B13

Sixty-four backpackers were asked the number of days since their latest backpacking trip. The number of days is given in **Table B26**:

# of days	1	2	3	4	5	6	7	8
Frequency	5	9	6	12	7	10	5	10

Table B26

33. Conduct an appropriate test to determine if the distribution is uniform.

a. The *p*-value is > 0.10. There is insufficient information to conclude that the distribution is not uniform.
b. The *p*-value is < 0.01. There is sufficient information to conclude the distribution is not uniform.
c. The *p*-value is between 0.01 and 0.10, but without alpha (α) there is not enough information
d. There is no such test that can be conducted.

34. Which of the following statements is true when using one-way ANOVA?

a. The populations from which the samples are selected have different distributions.
b. The sample sizes are large.
c. The test is to determine if the different groups have the same means.
d. There is a correlation between the factors of the experiment.

Practice Final Exam 2 Solutions

Solutions

1. b. parameter.

2. a.

3. c. seven

4. c. 0.02

5. c. none of the above

6. d. $\frac{100}{140}$

7. a. ≈ 0

8. b. The values for x are: {1, 2, 3,..., 14}

9. c. 0.9417.

10. d. binomial

11. d. 8.7

12. a. –1.96

13. a. 0.6321

14. d. 360

15. b. $N\left(72, \frac{72}{\sqrt{50}}\right)$

16. a. $\frac{3}{9}$

17. d.

18. b. 5.5

19. d. 6.92

20. a. 5

21. b. 0.8541

22. b. 0.2

23. a. –1.

24. c. matched pairs, dependent groups.

25. d. Reject H_0. There is sufficient evidence to conclude that there is a difference in the mean scores.

26. c. there is sufficient evidence to conclude that the proportion for males is higher than the proportion for females.

27. b. no

28. b. *p*-value is close to 1.

29. b. No

30. c. $\hat{y} = 79.96x - 0.0094$

31. d. We should not be estimating here.

32. a.

33. a. The *p*-value is > 0.10. There is insufficient information to conclude that the distribution is not uniform.

34. c. The test is to determine if the different groups have the same means.

APPENDIX C: DATA SETS

Lap Times

The following tables provide lap times from Terri Vogel's log book. Times are recorded in seconds for 2.5-mile laps completed in a series of races and practice runs.

	Lap 1	Lap 2	Lap 3	Lap 4	Lap 5	Lap 6	Lap 7
Race 1	135	130	131	132	130	131	133
Race 2	134	131	131	129	128	128	129
Race 3	129	128	127	127	130	127	129
Race 4	125	125	126	125	124	125	125
Race 5	133	132	132	132	131	130	132
Race 6	130	130	130	129	129	130	129
Race 7	132	131	133	131	134	134	131
Race 8	127	128	127	130	128	126	128
Race 9	132	130	127	128	126	127	124
Race 10	135	131	131	132	130	131	130
Race 11	132	131	132	131	130	129	129
Race 12	134	130	130	130	131	130	130
Race 13	128	127	128	128	128	129	128
Race 14	132	131	131	131	132	130	130
Race 15	136	129	129	129	129	129	129
Race 16	129	129	129	128	128	129	129
Race 17	134	131	132	131	132	132	132
Race 18	129	129	130	130	133	133	127
Race 19	130	129	129	129	129	129	128
Race 20	131	128	130	128	129	130	130

Table C1 Race Lap Times (in seconds)

	Lap 1	Lap 2	Lap 3	Lap 4	Lap 5	Lap 6	Lap 7
Practice 1	142	143	180	137	134	134	172
Practice 2	140	135	134	133	128	128	131
Practice 3	130	133	130	128	135	133	133
Practice 4	141	136	137	136	136	136	145

Table C2 Practice Lap Times (in seconds)

	Lap 1	Lap 2	Lap 3	Lap 4	Lap 5	Lap 6	Lap 7
Practice 5	140	138	136	137	135	134	134
Practice 6	142	142	139	138	129	129	127
Practice 7	139	137	135	135	137	134	135
Practice 8	143	136	134	133	134	133	132
Practice 9	135	134	133	133	132	132	133
Practice 10	131	130	128	129	127	128	127
Practice 11	143	139	139	138	138	137	138
Practice 12	132	133	131	129	128	127	126
Practice 13	149	144	144	139	138	138	137
Practice 14	133	132	137	133	134	130	131
Practice 15	138	136	133	133	132	131	131

Table C2 Practice Lap Times (in seconds)

Stock Prices

The following table lists initial public offering (IPO) stock prices for all 1999 stocks that at least doubled in value during the first day of trading.

$17.00	$23.00	$14.00	$16.00	$12.00	$26.00
$20.00	$22.00	$14.00	$15.00	$22.00	$18.00
$18.00	$21.00	$21.00	$19.00	$15.00	$21.00
$18.00	$17.00	$15.00	$25.00	$14.00	$30.00
$16.00	$10.00	$20.00	$12.00	$16.00	$17.44
$16.00	$14.00	$15.00	$20.00	$20.00	$16.00
$17.00	$16.00	$15.00	$15.00	$19.00	$48.00
$16.00	$18.00	$9.00	$18.00	$18.00	$20.00
$8.00	$20.00	$17.00	$14.00	$11.00	$16.00
$19.00	$15.00	$21.00	$12.00	$8.00	$16.00
$13.00	$14.00	$15.00	$14.00	$13.41	$28.00
$21.00	$17.00	$28.00	$17.00	$19.00	$16.00
$17.00	$19.00	$18.00	$17.00	$15.00	
$14.00	$21.00	$12.00	$18.00	$24.00	
$15.00	$23.00	$14.00	$16.00	$12.00	
$24.00	$20.00	$14.00	$14.00	$15.00	
$14.00	$19.00	$16.00	$38.00	$20.00	
$24.00	$16.00	$8.00	$18.00	$17.00	
$16.00	$15.00	$7.00	$19.00	$12.00	
$8.00	$23.00	$12.00	$18.00	$20.00	
$21.00	$34.00	$16.00	$26.00	$14.00	

Table C3 IPO Offer Prices

References

Data compiled by Jay R. Ritter of University of Florida using data from *Securities Data Co.* and *Bloomberg*.

APPENDIX D: GROUP AND PARTNER PROJECTS

Univariate Data

Student Learning Objectives

- The student will design and carry out a survey.
- The student will analyze and graphically display the results of the survey.

Instructions

As you complete each task below, check it off. Answer all questions in your summary.
____ Decide what data you are going to study.

Here are two examples, but you may **NOT** use them: number of M&M's per bag, number of pencils students have in their backpacks.

____ Are your data discrete or continuous? How do you know?
____ Decide how you are going to collect the data (for instance, buy 30 bags of M&M's; collect data from the World Wide Web).
____ Describe your sampling technique in detail. Use cluster, stratified, systematic, or simple random (using a random number generator) sampling. Do not use convenience sampling. Which method did you use? Why did you pick that method?
____ Conduct your survey. **Your data size must be at least 30.**
____ Summarize your data in a chart with columns showing **data value, frequency, relative frequency and cumulative relative frequency.**
Answer the following (rounded to two decimal places):

a. $\bar{x}$ = _____
b. s = _____
c. First quartile = _____
d. Median = _____
e. 70^{th} percentile = _____

____ What value is two standard deviations above the mean?
____ What value is 1.5 standard deviations below the mean?
____ Construct a histogram displaying your data.
____ In complete sentences, describe the shape of your graph.
____ Do you notice any potential outliers? If so, what values are they? Show your work in how you used the potential outlier formula to determine whether or not the values might be outliers.
____ Construct a box plot displaying your data.
____ Does the middle 50% of the data appear to be concentrated together or spread apart? Explain how you determined this.
____ Looking at both the histogram and the box plot, discuss the distribution of your data.

Assignment Checklist

You need to turn in the following typed and stapled packet, with pages in the following order:

____ **Cover sheet**: name, class time, and name of your study
____ **Summary page**: This should contain paragraphs written with complete sentences. It should include answers to all the questions above. It should also include statements describing the population under study, the sample, a parameter or parameters being studied, and the statistic or statistics produced.
____ **URL** for data, if your data are from the World Wide Web
____ **Chart of data, frequency, relative frequency, and cumulative relative frequency**
____ **Page(s) of graphs:** histogram and box plot

Continuous Distributions and Central Limit Theorem

Student Learning Objectives

- The student will collect a sample of continuous data.
- The student will attempt to fit the data sample to various distribution models.
- The student will validate the central limit theorem.

Instructions

As you complete each task below, check it off. Answer all questions in your summary.

Part I: Sampling

____ Decide what **continuous** data you are going to study. (Here are two examples, but you may NOT use them: the amount of money a student spent on college supplies this term, or the length of time distance telephone call lasts.)
____ Describe your sampling technique in detail. Use cluster, stratified, systematic, or simple random (using a random number generator) sampling. Do not use convenience sampling. What method did you use? Why did you pick that method?
____ Conduct your survey. Gather **at least 150 pieces of continuous, quantitative data**.
____ Define (in words) the random variable for your data. $X =$ _____
____ Create two lists of your data: (1) unordered data, (2) in order of smallest to largest.
____ Find the sample mean and the sample standard deviation (rounded to two decimal places).

a. $\bar{x} =$ _____
b. $s =$ _____

____ Construct a histogram of your data containing five to ten intervals of equal width. The histogram should be a representative display of your data. Label and scale it.

Part II: Possible Distributions

____ Suppose that X followed the following theoretical distributions. Set up each distribution using the appropriate information from your data.
____ Uniform: $X \sim U$ _____ Use the lowest and highest values as a and b.
____ Normal: $X \sim N$ _____ Use $\bar{x}$ to estimate for μ and s to estimate for σ.
____ **Must** your data fit one of the above distributions? Explain why or why not.
____ **Could** the data fit two or three of the previous distributions (at the same time)? Explain.
____ Calculate the value k(an X value) that is 1.75 standard deviations above the sample mean. $k =$ _____ (rounded to two decimal places) Note: $k = \bar{x} + (1.75)s$
____ Determine the relative frequencies (RF) rounded to four decimal places.

NOTE

$$RF = \frac{\text{frequency}}{\text{total number surveyed}}$$

a. $RF(X < k) =$ _____
b. $RF(X > k) =$ _____
c. $RF(X = k) =$ _____

NOTE

You should have one page for the uniform distribution, one page for the exponential distribution, and one page for the normal distribution.

____ State the distribution: $X \sim$ _____
____ Draw a graph for each of the three theoretical distributions. Label the axes and mark them appropriately.
____ Find the following theoretical probabilities (rounded to four decimal places).

a. $P(X < k) =$ _____
b. $P(X > k) =$ _____
c. $P(X = k) =$ _____

____ Compare the relative frequencies to the corresponding probabilities. Are the values close?
____ Does it appear that the data fit the distribution well? Justify your answer by comparing the probabilities to the relative frequencies, and the histograms to the theoretical graphs.

Part III: CLT Experiments

_____ From your original data (before ordering), use a random number generator to pick 40 samples of size five. For each sample, calculate the average.
_____ On a separate page, attached to the summary, include the 40 samples of size five, along with the 40 sample averages.
_____ List the 40 averages in order from smallest to largest.
_____ Define the random variable, $\bar{X}$, in words. $\bar{X}$ = _____
_____ State the approximate theoretical distribution of $\bar{X}$. $\bar{X} \sim$ _____
_____ Base this on the mean and standard deviation from your original data.
_____ Construct a histogram displaying your data. Use five to six intervals of equal width. Label and scale it.

Calculate the value $\bar{k}$ (an $\bar{X}$ value) that is 1.75 standard deviations above the sample mean. $\bar{k}$ = _____ (rounded to two decimal places)
Determine the relative frequencies (*RF*) rounded to four decimal places.

a. $RF(\bar{X} < \bar{k}) =$ _____
b. $RF(\bar{X} > \bar{k}) =$ _____
c. $RF(\bar{X} = \bar{k}) =$ _____

Find the following theoretical probabilities (rounded to four decimal places).

a. $P(\bar{X} < \bar{k}) =$ _____
b. $P(\bar{X} > \bar{k}) =$ _____
c. $P(\bar{X} = \bar{k}) =$ _____

_____ Draw the graph of the theoretical distribution of $\bar{X}$.
_____ Compare the relative frequencies to the probabilities. Are the values close?
_____ Does it appear that the data of averages fit the distribution of $\bar{X}$ well? Justify your answer by comparing the probabilities to the relative frequencies, and the histogram to the theoretical graph.
In three to five complete sentences for each, answer the following questions. Give thoughtful explanations.
_____ In summary, do your original data seem to fit the uniform, exponential, or normal distributions? Answer why or why not for each distribution. If the data do not fit any of those distributions, explain why.
_____ What happened to the shape and distribution when you averaged your data? **In theory,** what should have happened? In theory, would "it" always happen? Why or why not?
_____ Were the relative frequencies compared to the theoretical probabilities closer when comparing the X or $\bar{X}$ distributions? Explain your answer.

Assignment Checklist

You need to turn in the following typed and stapled packet, with pages in the following order:
____ **Cover sheet**: name, class time, and name of your study
____ **Summary pages**: These should contain several paragraphs written with complete sentences that describe the experiment, including what you studied and your sampling technique, as well as answers to all of the questions previously asked questions
____ **URL** for data, if your data are from the World Wide Web
____ **Pages, one for each theoretical distribution**, with the distribution stated, the graph, and the probability questions answered
____ **Pages of the data requested**
____ **All graphs required**

Hypothesis Testing-Article

Student Learning Objectives

- The student will identify a hypothesis testing problem in print.
- The student will conduct a survey to verify or dispute the results of the hypothesis test.
- The student will summarize the article, analysis, and conclusions in a report.

Instructions

As you complete each task, check it off. Answer all questions in your summary.
____**Find an article** in a newspaper, magazine, or on the internet which makes a claim about **ONE** population mean or **ONE** population proportion. The claim may be based upon a survey that the article was reporting on. Decide whether this claim is the null or alternate hypothesis.
____**Copy or print out the article** and include a copy in your project, along with the source.
____**State how you will collect your data.** (Convenience sampling is not acceptable.)
____**Conduct your survey. You must have more than 50 responses in your sample.** When you hand in your final project, attach the tally sheet or the packet of questionnaires that you used to collect data. Your data must be real.
____**State the statistics** that are a result of your data collection: sample size, sample mean, and sample standard deviation, OR sample size and number of successes.
____**Make two copies of the appropriate solution sheet.**
____**Record the hypothesis test** on the solution sheet, based on your experiment. **Do a DRAFT solution** first on one of the solution sheets and check it over carefully. Have a classmate check your solution to see if it is done correctly. Make your decision using a 5% level of significance. Include the 95% confidence interval on the solution sheet.
____**Create a graph that illustrates your data.** This may be a pie or bar graph or may be a histogram or box plot, depending on the nature of your data. Produce a graph that makes sense for your data and gives useful visual information about your data. You may need to look at several types of graphs before you decide which is the most appropriate for the type of data in your project.
____**Write your summary** (in complete sentences and paragraphs, with proper grammar and correct spelling) that describes the project. The summary **MUST** include:

a. Brief discussion of the article, including the source
b. Statement of the claim made in the article (one of the hypotheses).
c. Detailed description of how, where, and when you collected the data, including the sampling technique; did you use cluster, stratified, systematic, or simple random sampling (using a random number generator)? As previously mentioned, convenience sampling is not acceptable.
d. Conclusion about the article claim in light of your hypothesis test; this is the conclusion of your hypothesis test, stated in words, in the context of the situation in your project in sentence form, as if you were writing this conclusion for a non-statistician.
e. Sentence interpreting your confidence interval in the context of the situation in your project

Assignment Checklist

Turn in the following typed (12 point) and stapled packet for your final project:
____**Cover sheet** containing your name(s), class time, and the name of your study
____**Summary**, which includes all items listed on summary checklist
____**Solution sheet** neatly and completely filled out. The solution sheet does not need to be typed.
____**Graphic representation of your data**, created following the guidelines previously discussed; include only graphs which are appropriate and useful.

____**Raw data collected AND a table summarizing the sample data** (n, $\bar{x}$ and s; or x, n, and p', as appropriate for your hypotheses); the raw data does not need to be typed, but the summary does. Hand in the data as you collected it. (Either attach your tally sheet or an envelope containing your questionnaires.)

Bivariate Data, Linear Regression, and Univariate Data

Student Learning Objectives

- The students will collect a bivariate data sample through the use of appropriate sampling techniques.
- The student will attempt to fit the data to a linear model.
- The student will determine the appropriateness of linear fit of the model.
- The student will analyze and graph univariate data.

Instructions

1. As you complete each task below, check it off. Answer all questions in your introduction or summary.
2. Check your course calendar for intermediate and final due dates.
3. Graphs may be constructed by hand or by computer, unless your instructor informs you otherwise. All graphs must be neat and accurate.
4. All other responses must be done on the computer.
5. Neatness and quality of explanations are used to determine your final grade.

Part I: Bivariate Data

Introduction

____State the bivariate data your group is going to study.

Here are two examples, but you may **NOT** use them: height vs. weight and age vs. running distance.

____Describe your sampling technique in detail. Use cluster, stratified, systematic, or simple random sampling (using a random number generator) sampling. Convenience sampling is **NOT** acceptable.
____Conduct your survey. Your number of pairs must be at least 30.
____Print out a copy of your data.

Analysis

____On a separate sheet of paper construct a scatter plot of the data. Label and scale both axes.
____State the least squares line and the correlation coefficient.
____On your scatter plot, in a different color, construct the least squares line.
____Is the correlation coefficient significant? Explain and show how you determined this.
____Interpret the slope of the linear regression line in the context of the data in your project. Relate the explanation to your data, and quantify what the slope tells you.
____Does the regression line seem to fit the data? Why or why not? If the data does not seem to be linear, explain if any other model seems to fit the data better.
____Are there any outliers? If so, what are they? Show your work in how you used the potential outlier formula in the Linear Regression and Correlation chapter (since you have bivariate data) to determine whether or not any pairs might be outliers.

Part II: Univariate Data

In this section, you will use the data for **ONE** variable only. Pick the variable that is more interesting to analyze. For example: if your independent variable is sequential data such as year with 30 years and one piece of data per year, your x-values might be 1971, 1972, 1973, 1974, …, 2000. This would not be interesting to analyze. In that case, choose to use the dependent variable to analyze for this part of the project.
_____Summarize your data in a chart with columns showing data value, frequency, relative frequency, and cumulative relative frequency.
_____Answer the following question, rounded to two decimal places:

a. Sample mean = _____
b. Sample standard deviation = _____
c. First quartile = _____
d. Third quartile = _____
e. Median = _____
f. 70th percentile = _____
g. Value that is 2 standard deviations above the mean = _____
h. Value that is 1.5 standard deviations below the mean = _____

_____Construct a histogram displaying your data. Group your data into six to ten intervals of equal width. Pick regularly spaced intervals that make sense in relation to your data. For example, do NOT group data by age as 20-26,27-33,34-40,41-47,48-54,55-61 . . . Instead, maybe use age groups 19.5-24.5, 24.5-29.5, . . . or 19.5-29.5, 29.5-39.5, 39.5-49.5, . . .
_____In complete sentences, describe the shape of your histogram.
_____Are there any potential outliers? Which values are they? Show your work and calculations as to how you used the potential outlier formula in **Descriptive Statistics** (since you are now using univariate data) to determine which values might be outliers.
_____Construct a box plot of your data.
_____Does the middle 50% of your data appear to be concentrated together or spread out? Explain how you determined this.
_____Looking at both the histogram AND the box plot, discuss the distribution of your data. For example: how does the spread of the middle 50% of your data compare to the spread of the rest of the data represented in the box plot; how does this correspond to your description of the shape of the histogram; how does the graphical display show any outliers you may have found; does the histogram show any gaps in the data that are not visible in the box plot; are there any interesting features of your data that you should point out.

Due Dates

- Part I, Intro: _____ (keep a copy for your records)
- Part I, Analysis: _____ (keep a copy for your records)
- Entire Project, typed and stapled: _____

  ____ Cover sheet: names, class time, and name of your study

  ____ Part I: label the sections “Intro” and “Analysis.”
  ____ Part II:
  ____ Summary page containing several paragraphs written in complete sentences describing the experiment, including what you studied and how you collected your data. The summary page should also include answers to ALL the questions asked above.
  ____ All graphs requested in the project
  ____ All calculations requested to support questions in data
  ____ Description: what you learned by doing this project, what challenges you had, how you overcame the challenges

NOTE

Include answers to ALL questions asked, even if not explicitly repeated in the items above.

APPENDIX E: SOLUTION SHEETS

Hypothesis Testing with One Sample

Class Time: _____
Name: _____

a. H_0: _____
b. H_a: _____
c. In words, **CLEARLY** state what your random variable $\bar{X}$ or P' represents.
d. State the distribution to use for the test.
e. What is the test statistic?
f. What is the *p*-value? In one or two complete sentences, explain what the *p*-value means for this problem.
g. Use the previous information to sketch a picture of this situation. CLEARLY, label and scale the horizontal axis and shade the region(s) corresponding to the *p*-value.

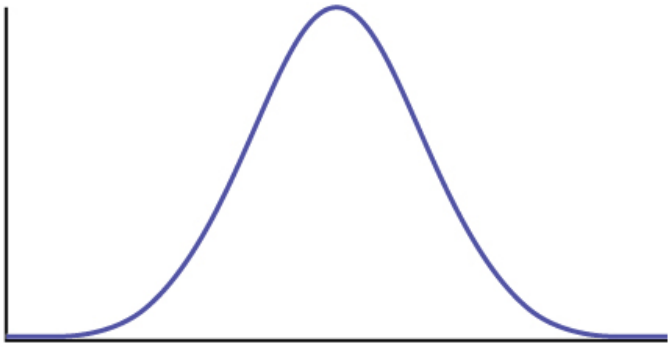

Figure E1

h. Indicate the correct decision ("reject" or "do not reject" the null hypothesis), the reason for it, and write an appropriate conclusion, using **complete sentences**.
    i. Alpha: _____
    ii. Decision: _____
    iii. Reason for decision: _____
    iv. Conclusion: _____
i. Construct a 95% confidence interval for the true mean or proportion. Include a sketch of the graph of the situation. Label the point estimate and the lower and upper bounds of the confidence interval.

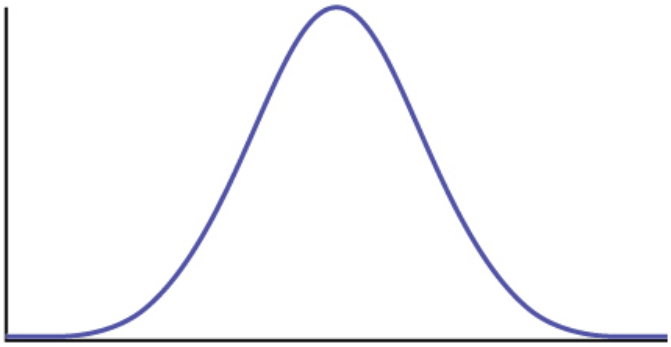

Figure E2

Hypothesis Testing with Two Samples

Class Time: _____
Name: _____

a. H_0: _____
b. H_a: _____
c. In words, **clearly** state what your random variable $\bar{X}_1 - \bar{X}_2$, $P'_1 - P'_2$ or $\bar{X}_d$ represents.
d. State the distribution to use for the test.
e. What is the test statistic?
f. What is the *p*-value? In one to two complete sentences, explain what the p-value means for this problem.
g. Use the previous information to sketch a picture of this situation. **CLEARLY** label and scale the horizontal axis and shade the region(s) corresponding to the *p*-value.

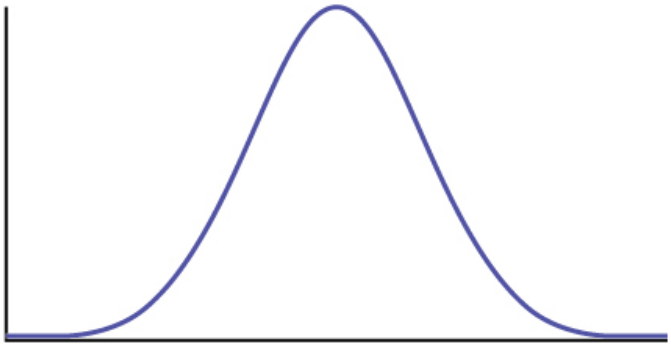

Figure E3

h. Indicate the correct decision ("reject" or "do not reject" the null hypothesis), the reason for it, and write an appropriate conclusion, using **complete sentences**.
    a. Alpha: _____
    b. Decision: _____
    c. Reason for decision: _____
    d. Conclusion: _____
i. In complete sentences, explain how you determined which distribution to use.

The Chi-Square Distribution

Class Time: _____
Name: _____

a. H_0: _____
b. H_a: _____
c. What are the degrees of freedom?
d. State the distribution to use for the test.
e. What is the test statistic?
f. What is the *p*-value? In one to two complete sentences, explain what the *p*-value means for this problem.
g. Use the previous information to sketch a picture of this situation. **Clearly** label and scale the horizontal axis and shade the region(s) corresponding to the *p*-value.

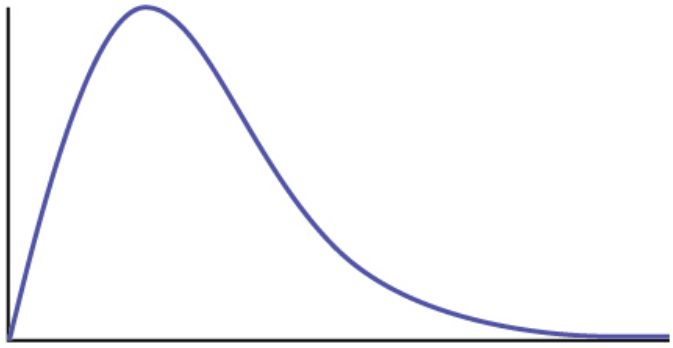

Figure E4

h. Indicate the correct decision ("reject" or "do not reject" the null hypothesis) and write appropriate conclusions, using **complete sentences.**
   i. Alpha: _____
   ii. Decision: _____
   iii. Reason for decision: _____
   iv. Conclusion: _____

F Distribution and One-Way ANOVA

Class Time: _____
Name: _____

a. H_0: _____
b. H_a: _____
c. $df(n)$ = _____ $df(d)$ = _____
d. State the distribution to use for the test.
e. What is the test statistic?
f. What is the *p*-value?
g. Use the previous information to sketch a picture of this situation. **Clearly** label and scale the horizontal axis and shade the region(s) corresponding to the *p*-value.

Figure E5

h. Indicate the correct decision (“reject” or “do not reject” the null hypothesis) and write appropriate conclusions, using **complete sentences**.
   a. Alpha: _____
   b. Decision: _____
   c. Reason for decision: _____
   d. Conclusion: _____

APPENDIX F: MATHEMATICAL PHRASES, SYMBOLS, AND FORMULAS

English Phrases Written Mathematically

When the English says:	Interpret this as:
X is at least 4.	$X \geq 4$
The minimum of X is 4.	$X \geq 4$
X is no less than 4.	$X \geq 4$
X is greater than or equal to 4.	$X \geq 4$
X is at most 4.	$X \leq 4$
The maximum of X is 4.	$X \leq 4$
X is no more than 4.	$X \leq 4$
X is less than or equal to 4.	$X \leq 4$
X does not exceed 4.	$X \leq 4$
X is greater than 4.	$X > 4$
X is more than 4.	$X > 4$
X exceeds 4.	$X > 4$
X is less than 4.	$X < 4$
There are fewer X than 4.	$X < 4$
X is 4.	$X = 4$
X is equal to 4.	$X = 4$
X is the same as 4.	$X = 4$
X is not 4.	$X \neq 4$
X is not equal to 4.	$X \neq 4$
X is not the same as 4.	$X \neq 4$
X is different than 4.	$X \neq 4$

Table F1

Formulas

Formula 1: Factorial

$$n! = n(n-1)(n-2)...(1)$$

$$0! = 1$$

Formula 2: Combinations

$$\binom{n}{r} = \frac{n!}{(n-r)!r!}$$

Formula 3: Binomial Distribution

$$X \sim B(n,\ p)$$

$$P(X = x) = \binom{n}{x} p^x q^{n-x}, \text{ for } x = 0,\ 1,\ 2,\ ...,\ n$$

Formula 4: Geometric Distribution

$$X \sim G(p)$$

$$P(X = x) = q^{x-1} p, \text{ for } x = 1,\ 2,\ 3,\ ...$$

Formula 5: Hypergeometric Distribution

$$X \sim H(r,\ b,\ n)$$

$$P(X = x) = \left(\frac{\binom{r}{x}\binom{b}{n-x}}{\binom{r+b}{n}} \right)$$

Formula 6: Poisson Distribution

$$X \sim P(\mu)$$

$$P(X = x) = \frac{\mu^x e^{-\mu}}{x!}$$

Formula 7: Uniform Distribution

$$X \sim U(a,\ b)$$

$$f(X) = \frac{1}{b-a},\ a < x < b$$

Formula 8: Exponential Distribution

$$X \sim Exp(m)$$

$$f(x) = me^{-mx} m > 0,\ x \geq 0$$

Formula 9: Normal Distribution

$$X \sim N(\mu,\ \sigma^2)$$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}},\ -\infty < x < \infty$$

Formula 10: Gamma Function

$$\Gamma(z) = \int_{\infty}^{0} x^{z-1} e^{-x} dx \quad z > 0$$

$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$

$\Gamma(m+1) = m!$ for m, a nonnegative integer

otherwise: $\Gamma(a+1) = a\Gamma(a)$

Formula 11: Student's *t*-distribution

$X \sim t_{df}$

$$f(x) = \frac{\left(1+\frac{x^2}{n}\right)^{\frac{-(n+1)}{2}} \Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}$$

$$X = \frac{Z}{\sqrt{\frac{Y}{n}}}$$

$Z \sim N(0, 1), Y \sim X^2_{df}$, n = degrees of freedom

Formula 12: Chi-Square Distribution

$X \sim X^2_{df}$

$f(x) = \frac{x^{\frac{n-2}{2}} e^{\frac{-x}{2}}}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}$, $x > 0$, n = positive integer and degrees of freedom

Formula 13: F Distribution

$X \sim F_{df(n),\, df(d)}$

$df(n)$ = degrees of freedom for the numerator

$df(d)$ = degrees of freedom for the denominator

$$f(x) = \frac{\Gamma(\frac{u+v}{2})}{\Gamma(\frac{u}{2})\Gamma(\frac{v}{2})}(\frac{u}{v})^{\frac{u}{2}} x^{(\frac{u}{2}-1)}[1+(\frac{u}{v})x^{-0.5(u+v)}]$$

$X = \frac{Y_u}{W_v}$, Y, W are chi-square

Symbols and Their Meanings

Chapter (1st used)	Symbol	Spoken	Meaning
Sampling and Data	$\sqrt{\ }$	The square root of	same
Sampling and Data	π	Pi	3.14159… (a specific number)
Descriptive Statistics	Q_1	Quartile one	the first quartile
Descriptive Statistics	Q_2	Quartile two	the second quartile
Descriptive Statistics	Q_3	Quartile three	the third quartile
Descriptive Statistics	*IQR*	interquartile range	$Q_3 - Q_1 = IQR$
Descriptive Statistics	$\bar{x}$	x-bar	sample mean

Table F2 Symbols and their Meanings

Chapter (1st used)	Symbol	Spoken	Meaning
Descriptive Statistics	μ	mu	population mean
Descriptive Statistics	s s_x sx	s	sample standard deviation
Descriptive Statistics	s^2 s_x^2	s squared	sample variance
Descriptive Statistics	σ σ_x σx	sigma	population standard deviation
Descriptive Statistics	σ^2 σ_x^2	sigma squared	population variance
Descriptive Statistics	Σ	capital sigma	sum
Probability Topics	$\{\}$	brackets	set notation
Probability Topics	S	S	sample space
Probability Topics	A	Event A	event A
Probability Topics	$P(A)$	probability of A	probability of A occurring
Probability Topics	$P(A\|B)$	probability of A given B	prob. of A occurring given B has occurred
Probability Topics	$P(A \text{ OR } B)$	prob. of A or B	prob. of A or B or both occurring
Probability Topics	$P(A \text{ AND } B)$	prob. of A and B	prob. of both A and B occurring (same time)
Probability Topics	A'	A-prime, complement of A	complement of A, not A
Probability Topics	$P(A')$	prob. of complement of A	same
Probability Topics	G_1	green on first pick	same
Probability Topics	$P(G_1)$	prob. of green on first pick	same
Discrete Random Variables	*PDF*	prob. distribution function	same
Discrete Random Variables	X	X	the random variable X
Discrete Random Variables	$X \sim$	the distribution of X	same
Discrete Random Variables	B	binomial distribution	same
Discrete Random Variables	G	geometric distribution	same
Discrete Random Variables	H	hypergeometric dist.	same
Discrete Random Variables	P	Poisson dist.	same
Discrete Random Variables	λ	Lambda	average of Poisson distribution
Discrete Random Variables	$\geq$	greater than or equal to	same
Discrete Random Variables	$\leq$	less than or equal to	same
Discrete Random Variables	=	equal to	same
Discrete Random Variables	≠	not equal to	same
Continuous Random Variables	$f(x)$	f of x	function of x
Continuous Random Variables	*pdf*	prob. density function	same
Continuous Random Variables	U	uniform distribution	same

Table F2 Symbols and their Meanings

Chapter (1st used)	Symbol	Spoken	Meaning
Continuous Random Variables	*Exp*	exponential distribution	same
Continuous Random Variables	*k*	*k*	critical value
Continuous Random Variables	*f*(*x*) =	*f* of *x* equals	same
Continuous Random Variables	*m*	*m*	decay rate (for exp. dist.)
The Normal Distribution	*N*	normal distribution	same
The Normal Distribution	*z*	*z*-score	same
The Normal Distribution	*Z*	standard normal dist.	same
The Central Limit Theorem	*CLT*	Central Limit Theorem	same
The Central Limit Theorem	$\bar{X}$	*X*-bar	the random variable *X*-bar
The Central Limit Theorem	μ_x	mean of *X*	the average of *X*
The Central Limit Theorem	$\mu_{\bar{x}}$	mean of *X*-bar	the average of *X*-bar
The Central Limit Theorem	σ_x	standard deviation of *X*	same
The Central Limit Theorem	$\sigma_{\bar{x}}$	standard deviation of *X*-bar	same
The Central Limit Theorem	ΣX	sum of *X*	same
The Central Limit Theorem	Σx	sum of *x*	same
Confidence Intervals	*CL*	confidence level	same
Confidence Intervals	*CI*	confidence interval	same
Confidence Intervals	*EBM*	error bound for a mean	same
Confidence Intervals	*EBP*	error bound for a proportion	same
Confidence Intervals	*t*	Student's *t*-distribution	same
Confidence Intervals	*df*	degrees of freedom	same
Confidence Intervals	$t_{\frac{\alpha}{2}}$	student t with *a*/2 area in right tail	same
Confidence Intervals	p' ; $\hat{p}$	*p*-prime; *p*-hat	sample proportion of success
Confidence Intervals	q' ; $\hat{q}$	*q*-prime; *q*-hat	sample proportion of failure
Hypothesis Testing	H_0	*H*-naught, *H*-sub 0	null hypothesis
Hypothesis Testing	H_a	*H*-*a*, *H*-sub *a*	alternate hypothesis
Hypothesis Testing	H_1	*H*-1, *H*-sub 1	alternate hypothesis
Hypothesis Testing	α	alpha	probability of Type I error
Hypothesis Testing	β	beta	probability of Type II error
Hypothesis Testing	$\overline{X1} - \overline{X2}$	*X*1-bar minus *X*2-bar	difference in sample means
Hypothesis Testing	$\mu_1 - \mu_2$	*mu*-1 minus *mu*-2	difference in population means

Table F2 Symbols and their Meanings

Chapter (1st used)	Symbol	Spoken	Meaning
Hypothesis Testing	$P'_1 - P'_2$	*P*1-prime minus *P*2-prime	difference in sample proportions
Hypothesis Testing	$p_1 - p_2$	*p*1 minus *p*2	difference in population proportions
Chi-Square Distribution	X^2	*Ky*-square	Chi-square
Chi-Square Distribution	O	Observed	Observed frequency
Chi-Square Distribution	E	Expected	Expected frequency
Linear Regression and Correlation	$y = a + bx$	*y* equals a plus *b*-*x*	equation of a line
Linear Regression and Correlation	$\hat{y}$	*y*-hat	estimated value of *y*
Linear Regression and Correlation	r	correlation coefficient	same
Linear Regression and Correlation	ε	error	same
Linear Regression and Correlation	*SSE*	Sum of Squared Errors	same
Linear Regression and Correlation	1.9*s*	1.9 times *s*	cut-off value for outliers
F-Distribution and ANOVA	*F*	*F*-ratio	*F*-ratio

Table F2 Symbols and their Meanings

APPENDIX G: NOTES FOR THE TI-83, 83+, 84, 84+ CALCULATORS

Quick Tips

Legend

-  represents a button press
- [ ] represents yellow command or green letter behind a key
- < > represents items on the screen

To adjust the contrast

Press  , then hold  to increase the contrast or  to decrease the contrast.

To capitalize letters and words

Press  to get one capital letter, or press  , then  to set all button presses to capital letters. You can return to the top-level button values by pressing  again.

To correct a mistake

If you hit a wrong button, just hit  and start again.

To write in scientific notation

Numbers in scientific notation are expressed on the TI-83, 83+, 84, and 84+ using E notation, such that...

- 4.321 E 4 = 4.321×10^{4}
- 4.321 E –4 = 4.321×10^{-4}

To transfer programs or equations from one calculator to another:

Both calculators: Insert your respective end of the link cable cable and press  , then [LINK].

Calculator receiving information:

1. Use the arrows to navigate to and select <RECEIVE>
2. Press  .

Calculator sending information:

1. Press appropriate number or letter.

2. Use up and down arrows to access the appropriate item.

3. Press ENTER to select item to transfer.
4. Press right arrow to navigate to and select <TRANSMIT>.

5. Press ENTER .

NOTE

ERROR 35 LINK generally means that the cables have not been inserted far enough.

Both calculators: Insert your respective end of the link cable cable Both calculators: press 2nd , then [QUIT] to exit when done.

Manipulating One-Variable Statistics

NOTE

These directions are for entering data with the built-in statistical program.

Data	Frequency
–2	10
–1	3
0	4
1	5
3	8

Table G1 Sample Data We are manipulating one-variable statistics.

To begin:

1. Turn on the calculator.

   ON

2. Access statistics mode.

   STAT

3. Select <4:ClrList> to clear data from lists, if desired.

   4 ENTER
   ,

4. Enter list [L1] to be cleared.

   2nd ENTER
   , [L1] ,

5. Display last instruction.

   2nd , [ENTRY]

6. Continue clearing remaining lists in the same fashion, if desired.

   ◀ , 2nd , [L2] , ENTER

7. Access statistics mode.

   STAT

8. Select <1:Edit . . .>

   ENTER

9. Enter data. Data values go into [L1]. (You may need to arrow over to [L1]).
   - Type in a data value and enter it. (For negative numbers, use the negate (-) key at the bottom of the keypad).

     (−) , 9 , ENTER

   - Continue in the same manner until all data values are entered.

10. In [L2], enter the frequencies for each data value in [L1].
    - Type in a frequency and enter it. (If a data value appears only once, the frequency is "1").

      4 , ENTER

    - Continue in the same manner until all data values are entered.

11. Access statistics mode.

    STAT

12. Navigate to <CALC>.
13. Access <1:1-var Stats>.

    ENTER

14. Indicate that the data is in [L1]...

    2nd , [L1] , ,

15. ...and indicate that the frequencies are in [L2].

    2nd , [L2] , ENTER

16. The statistics should be displayed. You may arrow down to get remaining statistics. Repeat as necessary.

Drawing Histograms

NOTE

We will assume that the data is already entered.

We will construct two histograms with the built-in STATPLOT application. The first way will use the default ZOOM. The second way will involve customizing a new graph.

1. Access graphing mode.

   2nd , [STAT PLOT]
2. Select <1:plot 1> to access plotting - first graph.

   ENTER
3. Use the arrows navigate go to <ON> to turn on Plot 1.

   ENTER <ON> ,
4. Use the arrows to go to the histogram picture and select the histogram.

   ENTER
5. Use the arrows to navigate to <Xlist>.
6. If "L1" is not selected, select it.

   2nd , [L1] , ENTER
7. Use the arrows to navigate to <Freq>.
8. Assign the frequencies to [L2].

   2nd , [L2] , ENTER
9. Go back to access other graphs.

   2nd , [STAT PLOT]
10. Use the arrows to turn off the remaining plots.
11. **Be sure to deselect or clear all equations before graphing.**

To deselect equations:

1. Access the list of equations.

   Y=
2. Select each equal sign (=).

   ▼ ▶ ENTER
3. Continue, until all equations are deselected.

To clear equations:

1. Access the list of equations.

   Y=
2. Use the arrow keys to navigate to the right of each equal sign (=) and clear them.

   ▼ ▶ CLEAR
3. Repeat until all equations are deleted.

To draw default histogram:

1. Access the ZOOM menu.

   ZOOM

2. Select <9:ZoomStat>.

   9

3. The histogram will show with a window automatically set.

To draw custom histogram:

1. Access window mode to set the graph parameters.

   WINDOW

2. - $X_{\min} = -2.5$
   - $X_{\max} = 3.5$
   - $X_{scl} = 1$ (width of bars)
   - $Y_{\min} = 0$
   - $Y_{\max} = 10$
   - $Y_{scl} = 1$ (spacing of tick marks on y-axis)
   - $X_{res} = 1$

3. Access graphing mode to see the histogram.

   GRAPH

To draw box plots:

1. Access graphing mode.

   2nd , [STAT PLOT]

2. Select <1:Plot 1> to access the first graph.

   ENTER

3. Use the arrows to select <ON> and turn on Plot 1.

   ENTER

4. Use the arrows to select the box plot picture and enable it.

   ENTER

5. Use the arrows to navigate to <Xlist>.
6. If "L1" is not selected, select it.

   2nd ENTER , [L1] ,

7. Use the arrows to navigate to <Freq>.
8. Indicate that the frequencies are in [L2].

   2nd ENTER , [L2] ,

9. Go back to access other graphs.

   2nd , [STAT PLOT]

10. **Be sure to deselect or clear all equations before graphing** using the method mentioned above.
11. View the box plot.

    GRAPH , [STAT PLOT]

Linear Regression

Sample Data

The following data is real. The percent of declared ethnic minority students at De Anza College for selected years from 1970–1995 was:

Year	Student Ethnic Minority Percentage
1970	14.13
1973	12.27
1976	14.08
1979	18.16
1982	27.64
1983	28.72
1986	31.86
1989	33.14
1992	45.37
1995	53.1

Table G2 The independent variable is "Year," while the independent variable is "Student Ethnic Minority Percent."

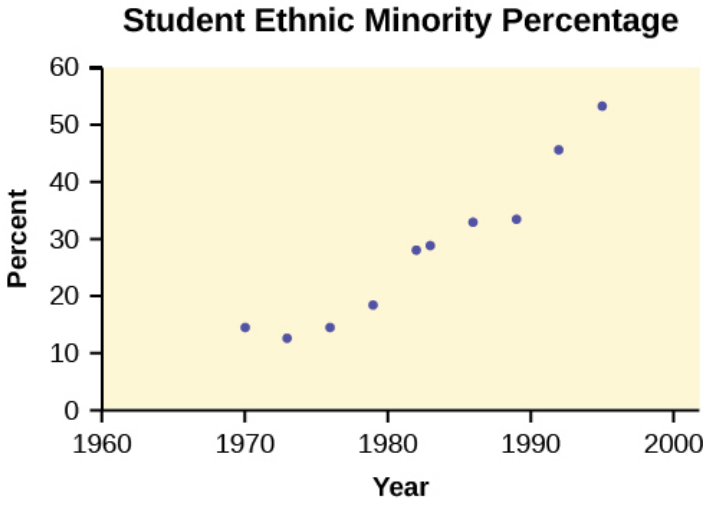

Figure G1 Student Ethnic Minority Percentage By hand, verify the scatterplot above.

NOTE

The TI-83 has a built-in linear regression feature, which allows the data to be edited.The *x*-values will be in [L1]; the *y*-values in [L2].

To enter data and do linear regression:

1. ON Turns calculator on.

   ON

2. Before accessing this program, be sure to turn off all plots.
   - Access graphing mode.

     2nd , [STAT PLOT]
   - Turn off all plots.

     4 ENTER ,
3. Round to three decimal places. To do so:
   - Access the mode menu.

     MODE , [STAT PLOT]
   - Navigate to <Float> and then to the right to <3>.

     ▼ ▶
   - All numbers will be rounded to three decimal places until changed.

     ENTER
4. Enter statistics mode and clear lists [L1] and [L2], as describe previously.

   STAT 4 ,
5. Enter editing mode to insert values for *x* and *y*.

   STAT ENTER ,

   ENTER
6. Enter each value. Press to continue.

To display the correlation coefficient:

1. Access the catalog.

   2nd , [CATALOG]
2. Arrow down and select <DiagnosticOn>

   ▼ ENTER ENTER ... , ,
3. r and r^2 will be displayed during regression calculations.

4. Access linear regression.

   STAT ▶

5. Select the form of $y = a + bx$.

   8 ENTER

   ,

The display will show:

LinReg

- $y = a + bx$
- $a = -3176.909$
- $b = 1.617$
- $r = 2\ 0.924$
- $r = 0.961$

This means the Line of Best Fit (Least Squares Line) is:

- $y = -3176.909 + 1.617x$
- Percent = –3176.909 + 1.617 (year #)

The correlation coefficient $r = 0.961$

To see the scatter plot:

1. Access graphing mode.

   2nd , [STAT PLOT]

2. Select <1:plot 1> To access plotting - first graph.

   ENTER

3. Navigate and select <ON> to turn on Plot 1.

   ENTER

   <ON>

4. Navigate to the first picture.
5. Select the scatter plot.

   ENTER

6. Navigate to <Xlist>.

7. If [L1] is not selected, press 2nd , [L1] to select it.
8. Confirm that the data values are in [L1].

   ENTER

   <ON>

9. Navigate to <Ylist>.
10. Select that the frequencies are in [L2].

    2nd ENTER , [L2] ,

11. Go back to access other graphs.

    2nd , [STAT PLOT]

12. Use the arrows to turn off the remaining plots.
13. Access window mode to set the graph parameters.

    WINDOW

    - $X_{\min} = 1970$
    - $X_{\max} = 2000$
    - $X_{scl} = 10$ (spacing of tick marks on x-axis)
    - $Y_{\min} = -0.05$
    - $Y_{\max} = 60$
    - $Y_{scl} = 10$ (spacing of tick marks on y-axis)
    - $X_{res} = 1$

14. Be sure to deselect or clear all equations before graphing, using the instructions above.

    GRAPH

15. Press the graph button to see the scatter plot.

To see the regression graph:

1. Access the equation menu. The regression equation will be put into Y1.

   Y=

2. Access the vars menu and navigate to <5: Statistics>.

   VARS 5 ,

3. Navigate to <EQ>.
4. <1: RegEQ> contains the regression equation which will be entered in Y1.

   ENTER

5. Press the graphing mode button. The regression line will be superimposed over the scatter plot.

   GRAPH

To see the residuals and use them to calculate the critical point for an outlier:

1. Access the list. RESID will be an item on the menu. Navigate to it.

   2nd , [LIST], <RESID>

2. Confirm twice to view the list of residuals. Use the arrows to select them.

   ENTER ENTER ,

3. The critical point for an outlier is: $1.9V\frac{\text{SSE}}{n-2}$ where:

    - n = number of pairs of data

- SSE = sum of the squared errors
- $\sum(\text{residual}^2)$

4. Store the residuals in [L3].

   STO▶ , 2nd , [L3] , ENTER

5. Calculate the $\frac{(\text{residual})^2}{n-2}$. Note that $n-2=8$

   2nd , [L3] , x² , ÷ , 8

6. Store this value in [L4].

   STO▶ , 2nd , [L4] , ENTER

7. Calculate the critical value using the equation above.

   1 , . , 9 , × , 2nd , [V] , 2nd , [LIST] , ▶ , ▶ , 5 , 2nd , [L4] ,) ,) , ENTER

8. Verify that the calculator displays: 7.642669563. This is the critical value.
9. Compare the absolute value of each residual value in [L3] to 7.64. If the absolute value is greater than 7.64, then the (x, y) corresponding point is an outlier. In this case, none of the points is an outlier.

To obtain estimates of *y* for various *x*-values:

There are various ways to determine estimates for "*y*." One way is to substitute values for "*x*" in the equation. Another way is to use the TRACE on the graph of the regression line.

TI-83, 83+, 84, 84+ instructions for distributions and tests

Distributions

Access DISTR (for "Distributions").

For technical assistance, visit the Texas Instruments website at **http://www.ti.com (http://www.ti.com)** and enter your calculator model into the "search" box.

Binomial Distribution

- binompdf(*n*,*p*,*x*) corresponds to $P(X = x)$
- binomcdf(*n*,*p*,*x*) corresponds to $P(X \le x)$
- To see a list of all probabilities for *x*: 0, 1, . . . , *n*, leave off the "*x*" parameter.

Poisson Distribution

- poissonpdf(λ,*x*) corresponds to $P(X = x)$
- poissoncdf(λ,*x*) corresponds to $P(X \le x)$

Continuous Distributions (general)

- $-\infty$ uses the value –1EE99 for left bound
- $+\infty$ uses the value 1EE99 for right bound

Normal Distribution

- normalpdf(*x*,*μ*,*σ*) yields a probability density function value (only useful to plot the normal curve, in which case "*x*" is the variable)
- normalcdf(left bound, right bound, *μ*, *σ*) corresponds to P(left bound < X < right bound)

- `normalcdf(left bound, right bound)` corresponds to $P(\text{left bound} < Z < \text{right bound})$ – standard normal
- `invNorm(p,μ,σ)` yields the critical value, k: $P(X < k) = p$
- `invNorm(p)` yields the critical value, k: $P(Z < k) = p$ for the standard normal

Student's *t*-Distribution

- `tpdf(x,df)` yields the probability density function value (only useful to plot the student-*t* curve, in which case "`x`" is the variable)
- `tcdf(left bound, right bound, df)` corresponds to $P(\text{left bound} < t < \text{right bound})$

Chi-square Distribution

- X^2`pdf(x,df)` yields the probability density function value (only useful to plot the chi^2 curve, in which case "`x`" is the variable)
- X^2`cdf(left bound, right bound, df)` corresponds to $P(\text{left bound} < X^2 < \text{right bound})$

F Distribution

- `Fpdf(x,dfnum,dfdenom)` yields the probability density function value (only useful to plot the *F* curve, in which case "`x`" is the variable)
- `Fcdf(left bound,right bound,dfnum,dfdenom)` corresponds to $P(\text{left bound} < F < \text{right bound})$

Tests and Confidence Intervals

Access `STAT` and `TESTS`.

For the confidence intervals and hypothesis tests, you may enter the data into the appropriate lists and press `DATA` to have the calculator find the sample means and standard deviations. Or, you may enter the sample means and sample standard deviations directly by pressing `STAT` once in the appropriate tests.

Confidence Intervals

- `ZInterval` is the confidence interval for mean when σ is known.
- `TInterval` is the confidence interval for mean when σ is unknown; *s* estimates σ.
- `1-PropZInt` is the confidence interval for proportion.

NOTE

The confidence levels should be given as percents (ex. enter "`95`" or "`.95`" for a 95% confidence level).

Hypothesis Tests

- `Z-Test` is the hypothesis test for single mean when σ is known.
- `T-Test` is the hypothesis test for single mean when σ is unknown; *s* estimates σ.
- `2-SampZTest` is the hypothesis test for two independent means when both σ's are known.
- `2-SampTTest` is the hypothesis test for two independent means when both σ's are unknown.
- `1-PropZTest` is the hypothesis test for single proportion.
- `2-PropZTest` is the hypothesis test for two proportions.
- X^2`-Test` is the hypothesis test for independence.
- X^2`GOF-Test` is the hypothesis test for goodness-of-fit (TI-84+ only).
- `LinRegTTEST` is the hypothesis test for Linear Regression (TI-84+ only).

NOTE

Input the null hypothesis value in the row below "`Inpt`." For a test of a single mean, "`μø`" represents the null hypothesis. For a test of a single proportion, "`pø`" represents the null hypothesis. Enter the alternate hypothesis on the bottom row.

APPENDIX H: TABLES

The module contains links to government site tables used in statistics.

NOTE

When you are finished with the table link, use the back button on your browser to return here.

Tables (NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/, January 3, 2009)

- **Student t table (http://www.itl.nist.gov/div898/handbook/eda/section3/eda3672.htm)**
- **Normal table (http://www.itl.nist.gov/div898/handbook/eda/section3/eda3671.htm)**
- **Chi-Square table (http://www.itl.nist.gov/div898/handbook/eda/section3/eda3674.htm)**
- **F-table (http://www.itl.nist.gov/div898/handbook/eda/section3/eda3673.htm)**
- All **four tables (http://www.itl.nist.gov/div898/handbook/eda/section3/eda367.htm)** can be accessed by going to

95% Critical Values of the Sample Correlation Coefficient Table

- **95% Critical Values of the Sample Correlation Coefficient**

INDEX

Symbols

A

B

C

D

E

F

G

H

I

T

U

V

Z